SAS 统计分析与应用实例

刘　荣　编著

電子工業出版社
Publishing House of Electronics Industry
北京 • BEIJING

内 容 简 介

SAS（Statistical Analysis System，统计分析系统）作为当今国际最著名的数据分析软件系统，功能强大、应用广泛。

本书基于 SAS 9.2 软件，结合编程和菜单操作以实例为载体介绍 SAS 系统。主要内容有：SAS 系统简介、SAS 编程简介、SAS 菜单操作、定量数据描述性统计分析、参数估计与假设检验、方差分析、相关与回归分析、列联表分析、非参数检验、主成分分析与因子分析、典型相关分析、聚类分析、判别分析。

本书内容丰富、结构清晰，采取从原理到实践的方式。随书附赠的光盘中收录了本书所有例题、实验、上机练习的数据与程序，并含长达 8 小时的全程实录讲解视频。

本书的读者对象是高等院校各专业学习应用统计的大学生、研究生，以及企事业单位数据分析工作者。

图书在版编目（CIP）数据

SAS 统计分析与应用实例/刘荣编著. —北京：电子工业出版社，2013.10
ISBN 978-7-121-21537-7

Ⅰ. ①S… Ⅱ. ①刘… Ⅲ. ①统计分析－应用软件 Ⅳ. ①C819

中国版本图书馆 CIP 数据核字（2013）第 224599 号

策划编辑：陈韦凯
责任编辑：王凌燕
印　　刷：
装　　订：
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编 100036
开　　本：787×1 092　1/16　印张：17.75　字数：454.4 千字
版　　次：2013 年 10 月第 1 版
印　　次：2019 年 1 月第 2 次印刷
定　　价：49.00 元（含 DVD 光盘 1 张）

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：chenwk@phei.com.cn。

前　言

SAS（Statistical Analysis System，统计分析系统）作为当今国际最著名的数据分析软件系统，被广泛应用于金融、保险、医疗卫生、教育和科研等各行业与领域。

本书的编排遵循从理论到实践的原则，首先介绍基本统计原理，然后结合实例应用SAS9.2 的编程和菜单操作的方式完成各项统计分析，并结合例题实际背景对分析结果进行详细阐述。

以下介绍本书的基本结构：本书共 13 章，前 3 章为基础知识介绍，后 10 章为统计分析实验。

第 1 章为 SAS 系统简介，介绍了 SAS 系统的应用范围、主要模块、界面操作、文件管理及 SAS 逻辑库与数据集的部分操作。

第 2 章为 SAS 编程简介，主要介绍了 SAS 数据步中的输入、赋值、控制语句，SAS 程序步中的 VAR、BY 等语句，并引入 ODS 输出系统和 SAS 宏。

第 3 章为 SAS 菜单操作，引入 SAS/ASSIST、SAS/INSIGHT 和 SAS/Analyst 三个菜单操作模块。

第 4 章为定量数据描述性统计分析，介绍了连续型数据描述性统计分析方法，以及统计图形的绘制。

第 5 章为参数估计与假设检验，介绍了单样本均值和方差的区间估计、独立和配对样本 T 检验及正态分布拟合检验。

第 6 章为方差分析，主要介绍单因素、多因素方差分析（包括区组设计、析因设计和拉丁方设计）和协方差分析。

第 7 章为相关与回归分析，介绍了相关分析、线性回归和非线性回归。

第 8 章为列联表分析，介绍了列联表的编制及其拟合优度检验、独立性检验、一致性检验、趋势检验和计算属性关联度。

第 9 章为非参数检验，介绍了单样本、两独立（配对）样本和多个独立样本的位置检验。

第 10 章到第 13 章为多元回归分析部分，主要包括主成分与因子分析、典型相关分析、判别与聚类分析。

本书内容丰富、层次清晰。在程序中添加了详细的注释；菜单操作过程介绍清晰。随书附赠的光盘中分章收录了本书所有的例题、实验、上机练习的数据和程序；并包含 8 小时随书全程实录，力求让读者形象、直观地掌握 SAS 系统的一般统计分析方法。

本书的读者对象是高等院校各专业学习应用统计的大学生、研究生，以及企事业单位数据分析工作者。

本书主要由刘荣编写，同时，参与本书编写工作的还有张玉兰、孙明、唐伟、王杨、顾辉、李成、陈杰、张霁芬、张计、陈军、张强、杨明、李建、李兵等人。

因作者水平有限，书中错误、纰漏之处难免，欢迎广大读者批评指正。

编 著 者

目　　录

第 1 章　SAS 系统简介

SAS（Statistical Analysis System）系统于 1976 年由SAS 软件研究所（SAS Institute Inc.）研制推出。历经多年发展，最新版本 SAS9.3 于 2011 年 7 月问世，它作为国际公认的著名数据统计分析软件系统之一，受到许多国家和地区的机构青睐。本章将简介 SAS 系统主要模块、界面操作和文件管理，并在引入 SAS 逻辑库和数据库后介绍新建逻辑库、数据集的方法，及数据集的排序、连接、合并及与外部数据文件相互转换等操作的实现，在章末将重点介绍 SAS 帮助系统的使用，“授人以鱼，不如授人以渔”，希望读者能举一反三，将 SAS 软件灵活运用于学习和工作实践中。

1.1　SAS 系统概述

SAS 系统将数据管理和统计分析融为一体，能够在不同的操作系统（如 UNIX、MS-DOS、VMS 等）和不同的机器类型下运行，系统具备完备的数据存取、数据管理、数据分析和数据展示的功能。在 Windows 版的 SAS 运行环境下，用户不仅能够以灵活的编程方式，还可选择操作简单的菜单方式进行各种统计分析。目前，SAS 系统被广泛应用于金融、医疗卫生、生产、运输、通信、科研和教育等领域。它运用统计分析、时间序列分析、运筹决策等科学的方法进行质量控制、财务管理、生产优化、风险管理、市场调查和预测等业务，并可将各种数据以灵活的报表、图形和三维透视的形式直观地表现出来。

SAS 系统包含众多模块以完成不同任务，本书内容涉及的有：

- SAS/BASE（基础）——完成数据整理和初步统计分析。
- SAS/STAT（统计）——广泛的统计分析。
- SAS/ASSIST（面向任务的通用菜单驱动界面）——交互式菜单操作。
- SAS/GRAPH（图形）——提供了许多产生图形过程并支持众多图形设备。

SAS/BASE 是 SAS 系统的核心平台，提供了多个 SAS 过程，可以实现简单报表计算、生成报表、计算得分（标准化数据、数据排秩等）及排序等功能。

SAS/STAT 提供 SAS 系统用来解决实际问题的具体统计分析过程，主要包括方差分析、回归分析、属性数据分析、多变量分析、聚类分析、判别分析等。

SAS/GRAPH 具有强大的绘图功能，能够用于绘制二维曲线图、条形图、饼图、区域图等。

本书将基于 Win7 系统下的 SAS9.2 多国语言版的 SAS（English（ODBS））版本（该版本系统语言为英文，但支持一般的中文字符输入），首先概述 SAS 系统、简介 SAS 编程和菜单操作，然后介绍具体统计方法基本原理及其在 SAS 系统中的实现。

注意：建议读者先快速浏览通读本书，对介绍的基本内容形成大致印象，在遇到实际问题时再具体参考相应章节。若时间充裕，也可以通过观看随书附赠的光盘中的教学视频练习本书

介绍的案例来熟练掌握 SAS 软件。

1.2 SAS 界面操作与文件管理

下面介绍 SAS 系统的启动和退出、界面特性及文件管理，完成以上操作前请用户将 SAS 软件安装在一台满足 SAS 系统配置的计算机中。

1.2.1 SAS 系统的启动与退出

SAS 系统的启动：

- 在系统的“开始”菜单列表中找到 SAS 系统文件夹，左键单击“SAS9.2（Additional Languages）”文件夹下的项目“SAS 9.2（English（DBCS））”即可启动系统，如图 1-1 所示。

图 1-1 启动 SAS 系统

- 若右键单击项目“SAS 9.2（English（DBCS））”，在下拉选项中选择“发送到”|“桌面快捷方式”，即可双击桌面上的图标启动 SAS 系统。

SAS 系统的退出：

- 选择菜单 File|Exit，在弹出的确认对话框中左键单击“确定”按钮。
- 单击系统主界面右上角按钮，在弹出的确认对话框中单击“确定”按钮。

1.2.2 SAS 系统界面简介

启动 SAS 系统将出现如图 1-2 所示的工作界面。它在一个主窗口内包含有若干个子窗口，

并有菜单栏、工具栏、状态栏等。下面介绍 SAS 系统界面的主要窗口：Editor 程序编辑窗口、Log 运行记录窗口、Output 输出记录窗口、Explorer 窗口和 Results 窗口。

- Editor 程序编辑窗口——主要用于编辑 SAS 源程序文件，操作时光标可在整个窗口随意移动，且支持 Windows 系统常规编辑操作方式，如剪切、复制、粘贴等。SAS9.2 的智能编辑功能可根据用户输入的不同的SAS程序部分显示出不同的颜色。若用户输入有误，对应的颜色不对，以警告错误的发生。
- Log 运行记录窗口——用于显示和记录 SAS 程序的运行情况，说明其运行成功或提示错误信息。当程序运行不成功，Log 运行记录窗口将分别用绿色字符和红色字符显示警告和错误信息。
- Output 输出记录窗口——分页显示 SAS 程序运行的文本型输出结果，可使用主界面菜单的 File|Save As 将其保存在磁盘中，文件扩展名为.lst，该类型文件可用文字处理软件如 Word、写字板、记事本等将其打开和编辑。SAS 程序运行的图形输出结果将由 Graphics 窗口显示，可选择菜单 File|Export as Image 将图形导出并保存在磁盘内，并可在“保存类型”下拉列表中选择图形文件的保存格式。
- Explorer 窗口——用于显示 SAS 逻辑库（SAS 系统命名的库名和磁盘某文件的关联）及 SAS 数据集。
- Results 窗口——用于显示 SAS 成功运行时程序输出结果的目录。

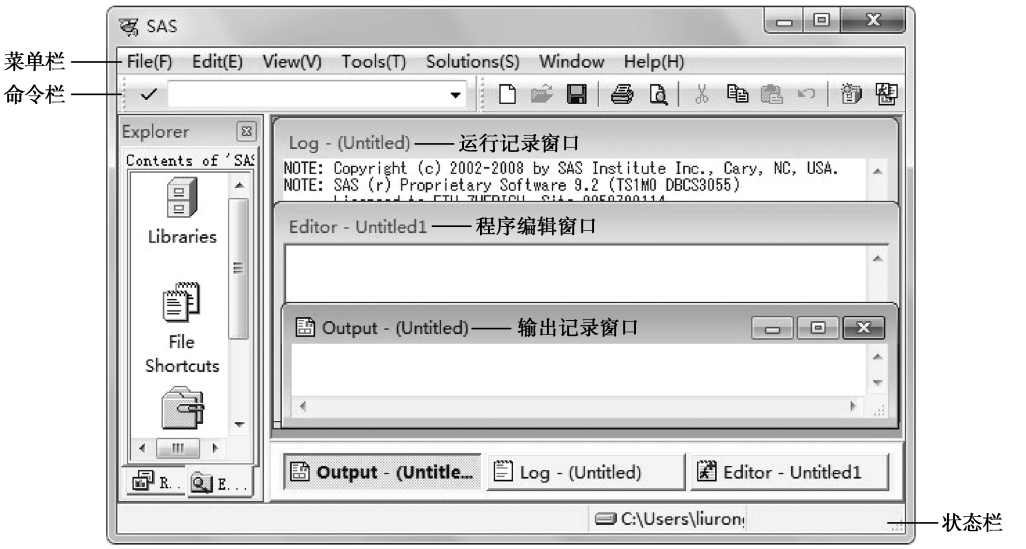


图 1-2　SAS 系统工作界面

可用左键单击窗口内的任意一处的方式切换到以上任一窗口。在使用时可根据需要直接单击窗口右上角的按钮关闭窗口，也可使用主菜单 View 的下拉菜单打开或切换窗口。

SAS 主窗口标题栏下是主菜单，注意 SAS 菜单随着光标出现在不同的窗口是动态变化的。其主要命令及功能实现如下：

- File（文件）——实现 SAS 文件的调入、保存、转换及打印等功能。
- Edit（编辑）——实现窗口的编辑（清空、剪切、复制、粘贴等）功能。
- View（浏览）——实现打开或切换到 SAS 的各个工作窗口的功能。

- Tools（工具）——提供对各种输出结果进行编辑的工具，如表格、图形、报告等，并支持进行系统环境和状态的设置（如修改界面字体等）。
- Run（运行）——提交程序，仅在当前窗口为 Editor 程序编辑窗口时有效。
- Solutions（解决方案）——SAS 图形界面模块操作窗口，如 SAS/ASSIST。
- Help（帮助）——提供 SAS 软件自带的帮助系统。

主菜单下是一个命令输入栏和图表工具栏。在命令栏中可以输入 SAS 的显示管理命令，如在此输入“WSAVE”则表示永久保存对 SAS 系统的设置。工具栏图标提供了常见任务的快捷操作方式，其功能解释如下：

- New——建立新的编辑窗口。
- Open——打开文件到编辑窗口。注意：用户指定一个文件调入到编辑窗口内，以后的存盘操作将自动存入此文件。
- Save——保存编辑窗口内容。注意：若此窗口已经与一个文件相联系，存盘功能将覆盖文件原有内容。
- Print——打印当前窗口内容。
- Print preview——打印预览。
- Cut——剪切选定文本。
- Copy——复制选定文本。
- Paster——粘贴。

注意：这些操作是对 Windows 剪贴板进行的，所以它不仅支持 SAS 编辑窗口内的复制和剪切操作，还可用来与其他 Windows 应用程序交换文本、数据等。

- Undo——撤销编辑操作。
- New Library——建立新的 SAS 逻辑库。
- SAS Explorer——打开 SAS 管理窗口查看、管理 SAS 的各个逻辑库及其中的文件。
- Submit——提交 Editor 程序编辑窗口中的程序。
- Clear All——清空当前窗口内容。
- Break——中断正在运行的 SAS 程序。
- Help——进入 SAS 的帮助界面。

1.2.3 SAS 数据集和逻辑库

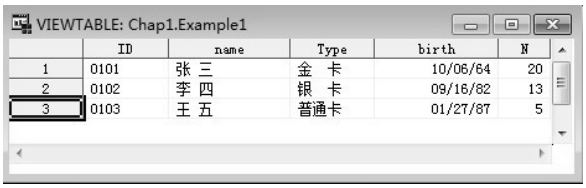
SAS 文件主要包括数据集（Database）文件、索引文件和 SAS 目录文件（Catalog）等。数据集是 SAS 使用和分析计算的原始数据来源，而正确合理地生成 SAS 数据集是数据分析的首要条件，因此数据集是 SAS 最重要的文件类型。SAS 目录文件主要用以保存各种不能表示成行列结构表格形式的数据，如系统设置、图像、声音等。

SAS 数据集可以看作由若干行和列组成的表格，数据集的每一行称为一个观测（Observation），每一列称为一个变量（Variable），变量可以取不同的类型值，如整数型、浮点值、时间值、字符串、货币值等。

如图 1-3 所示的数据集范例中包括了 3 条观测，代表了 3 个客户的情况；包含 5 个变量，分别为客户编号（ID）、姓名（name）、持卡类型（Type）、年龄（birth）和消费次数（N）。注意到该图中数据集名称为 Chap1.Example1，ID、name、Type、birth 和 N 为变量名。在 SAS 系统

中使用的数据集、变量名、逻辑库名等统称为“标识符”，SAS 系统对标识符有以下严格规定：

- SAS 标识符必须由英文字母、数字、下画线组成。
- 第一个字符必须是字母或下画线。
- 标识符中字母不区分大小写。
- 标识符的长度不宜过长。



	ID	name	Type	birth	N
1	0101	张 三	金　卡	10/06/64	20
2	0102	李 四	银　卡	09/16/82	13
3	0103	王 五	普通卡	01/27/87	5

图 1-3　数据集范例

由此可知 name、area、ABC、X2、_Nall_等都是合法标识符，area 和 AREA 为同一标识符，但 number-3（不能有减号）、a bit（不能有空格）、team*（不能有特殊字符）等却不合法。

SAS 数据集存储在被称为 SAS 逻辑库（Library）的文件集中。通俗地说，SAS 逻辑库是一个连接，将磁盘中存储的文件和 SAS 系统联系起来。SAS 逻辑库命名遵循上述 SAS 命名规则，可用编程和菜单操作的方式建立逻辑库。

编程建立逻辑库：使用 Libname 命令可以指定逻辑库，命名格式为：

```
Libname 逻辑库标记"文件夹路径";
```

例如，要建立指向已存在的文件目录“E:\data\chap1”的逻辑库 chap1，可在 Editor 程序编辑窗口输入以下语句：

```
Libname chap1 'E:\data\chap1';
```

再选择 Run|Submit 菜单或左键单击工具栏上的 图标提交程序完成操作。

菜单方式建立逻辑库：此方式操作过程如下：进入 Explorer 窗口，双击 Library 图标，再单击右键，选择 new 命令，在弹出的窗口（如图 1-4 所示）的 Name 栏中输入逻辑库名，在 Path 栏中输入路径或单击右侧的 Browse（浏览）按钮选择磁盘中的文件夹，选中右侧的 Enable at startup（启动时可用）使其永久有效，最后单击 OK 按钮完成。

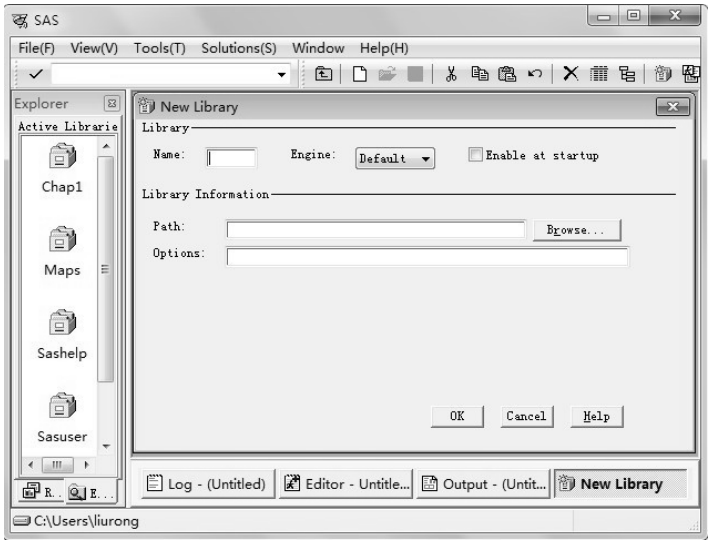


图 1-4　新建逻辑库

以上介绍了新建逻辑库的一般方法，接下来概述两种逻辑库类型：临时库和永久库，以及其对应的临时数据集和永久数据集的命名区别。

临时库和永久库：临时库仅有 WORK 逻辑库，它存储 SAS 临时文件，在每次启动 SAS 系统时自动生成，关闭 SAS 系统时库中的数据集被自动删除；永久库中存储 SAS 永久文件，SAS 预定义 Sasuser 和 Sashelp 两个永久逻辑库，其中 Sasuser 用以保存与个人设置有关的文件，Sashelp 逻辑库保存与 SAS 帮助系统、应用实例有关的文件。除此之外，用户可使用上述介绍的编程和菜单操作的方式建立 SAS 永久库。

临时数据集和永久数据集命名区别：临时数据集可以用“单水平名”，即只有数据集名，如 stu01，这种名字的数据集被保存在 WORK 逻辑库中，因此也可用 WORK.stu01 表示。永久数据集由两部分组成，前一部分为它的库名，后一部分为数据集名，中间用小数点连接，如放在 chap1 库（即“E:\data\chap1”子目录）中的数据集 ex1 需要用 chap1.ex1 表示。临时数据集在重启 SAS 系统时将会被自动删除，而永久数据集不会。

1.2.4 数据集文件操作

SAS 管理器（SAS Explorer）用来管理 SAS 逻辑库和文件，SAS 系统默认启动时在主界面左侧显示 SAS 管理器，若不慎将其关闭，可通过主菜单命令 View|Explorer 将其打开。以下介绍对逻辑库中数据文件的复制、删除、改名等操作。

- 文件复制：不同逻辑库之间的数据文件可以复制备份，操作为：左键双击逻辑库名，单击选择需要复制的数据文件，右键单击，在弹出的快捷菜单中选择 Copy（复制）命令实现复制，再左键双击打开目标逻辑库名，右键单击，在弹出的快捷菜单中选择 Paster（粘贴）命令完成备份。
- 文件删除：选择目标数据文件，右键单击，在弹出的快捷菜单中选择 Delete（删除）命令。
- 文件重命名：选择目标数据文件，右键单击，在弹出的快捷菜单中选择 Rename（更名）命令，再在弹出的对话框中输入改后的文件名，左键单击 OK 按钮保存设置并退出。

1.3 SAS 数据集整理

本章 1.2 节简单介绍了 SAS 的界面，并引入 SAS 逻辑库和数据集，详述了新建逻辑库的方法，以及数据集文件的复制、删除等操作。以下介绍 SAS 数据集的一般操作。

1.3.1 新建数据集

在 SAS 系统中可以使用编程和 Viewtable 表的方式新建 SAS 数据集，并对数据集进行修改、增删记录等操作。以下通过例 1-1 介绍这两种方法的具体使用。

例 1-1 请根据表 1-1 所示信息新建数据集 chap1.example1_1。

表 1-1　某公司客户信息表

编　　号 (ID)	姓　　名 (name)	持卡类型 (Type)	出生年月 (birth)	消费次数 (N)	登记时间 (Date)
0101	张 三	金卡	1964/10/06	20	2011/01/02
0102	李 四	银卡	1982/09/16	13	2011/04/03
0103	王 五	消费卡	1987/01/27	5	2011/05/06

方法一：编写如下程序（其在光盘中的存储路径为“proc\chap1\example1_1.sas”）。

```
libname chap1 'E:\data\chap1';      /*新建指向路径为“E：\data\chap1”的逻辑库 chap1*/
data chap1.example1_1;              /*新建永久 SAS 数据集 chap1.example1*/
input ID $1-4 name $5-11 Type$13-19 birth YYMMDD8. +1 N;     /*定义变量输入格式*/
cards;
0101 张 三   金   卡 64/10/06 20
0102 李 四   银   卡 82/09/16 13
0103 王 五   普通卡 87/01/27 5
;                                   /*输入数据*/
proc print;
format birth YYMMDD8.;              /*设置变量 birth 的输出格式*/
run;
```

选择 Run|Submit 命令提交程序，在 Output 输出记录窗口显示如图 1-5 所示，且在新建的逻辑库 chap1 中出现了数据集 example1。

图 1-5　结果输出

注意：日期格式的变量在 SAS 中默认保存为从 1960 年 1 月 1 日至某日期的天数，如 1964 年 10 月 6 号，默认输出为 1740。在打印输出日期格式的变量时，为了得到特定格式，务必使用 format 语句指定输出格式。而 SAS 数据集 chap1.example1，变量 birth 也将显示天数，此时右键单击变量名，在快捷菜单中选择 Column Attribute（变量属性），单击弹出的对话框中的 format 后的 按钮，在此可以选择设置变量的输出格式。

方法二：Viewtable 表新建数据集。

步骤一：打开新表

选择菜单 Tools|Table Editor 打开一个新表，如图 1-6 所示。

步骤二：定义变量

右键单击 A 列，选择 Column Attributes（变量属性）命令，弹出如图 1-7 所示对话框，在 Name 栏中输入变量名 ID，Type（类型）为默认 Character（字符型），Length（长度）为 8，Format（输出格式）和 Informat（输入格式）为$8.（默认设置）。左键单击 Apply（应用）按钮，再单击 Close（关闭）按钮；或者直接单击 Close 按钮完成第一列姓名变量的属性设置。

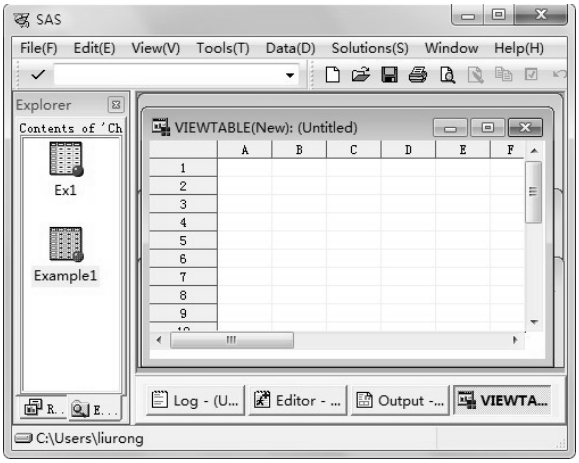


图 1-6　Viewtable 表视窗

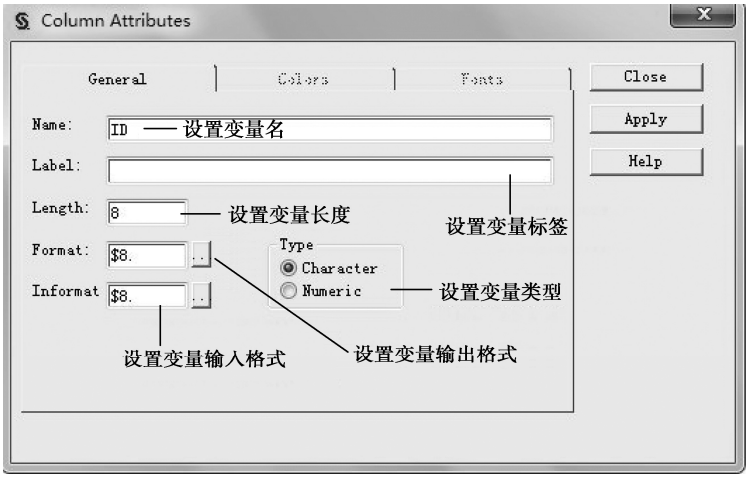


图 1-7　设置变量属性

重复以上操作，定义变量 Name、Type 时变量类型（Type）选择 Character，输出格式（Format）和输入格式（Informat）选择默认值$8.；定义变量 birth 时变量类型（Type）选择 Numeric，Format（输出格式）和 Informat（输入格式）选择 yy/mm/dd；定义变量 N 时变量类型（Type）选择 Numeric，Format（输出格式）和 Informat（输入格式）选择默认值$8.。

说明：用户根据变量性质选择设置变量类型（Type）为字符型（Character）和数值型（Numeric）。Informat 和 Format 分别代表数据的输入格式和输出格式，可单击格式右侧的“…”按钮设置。在窗口的 Label 栏中可以输入变量标签。若不定义变量标签，则系统默认变量名为标签名。打开数据集时，Viewtable 表头为标签名，可通过主菜单 View|Columns Labels 和 View|Columns Names 进行切换。

步骤三：输入数据、保存数据集

直接在表中输入相应数据，然后直接单击工具栏上的保存图标，或者选择菜单 File|Save，在弹出的对话框中左键单击选择需要存入的库名 chap1，在 Member Name（对象名称）框中输入数据集名 example1，单击 Save（保存）按钮保存设置。

若要修改已有数据集的变量标签，可选择菜单 Editor|Edit Mode（编辑模式）更改系统默认

的 Browse（浏览）模式。

若要更改已存在的数据集变量名，可使用 RENAME 语句用编程的方式实现，它的一般使用格式如下：

```
RENAME OLD1=NEW1 OLD2=NEW2 OLD3=NEW3;
```

注意：建议读者在观测和变量不多，数据形式不复杂的情形下选用编程的方式新建数据集，否则可选择便捷的 Viewtable 表建立 SAS 数据集。若是海量数据，则采用从外部数据库将数据导入 SAS 系统进行计算的方式。

1.3.2　在数据集中增加、筛选变量和观测

在 DATA 步中可以直接利用新建变量语句来产生新变量，筛选变量可使用 KEEP/DROP 语句，KEEP 语句用来指定需要保留的变量；DROP 语句用来指定被删除的变量。注意在一个程序里，不能同时使用 DROP/KEEP 语句。KEEP/DROP 语句的语句格式如下（注意：变量间用空格分开）：

- KEEP 语句格式 1：KEEP 变量 1 变量 2…变量 N;
- KEEP 语句格式 2：DATA=数据集名称（KEEP=变量 1 变量 2…变量 N）;
- DROP 语句格式 1：DROP 变量 1 变量 2…变量 N;
- DROP 语句格式 2：DATA=数据集名称（DROP=变量 1 变量 2…变量 N）;

在 SAS 数据集中可使用 IF-THEN 语句实现筛选观测。该语句有以下两种使用格式：

格式一：IF 条件表达式 THEN;

格式二：IF 条件表达式 THEN SAS 语句;

<ELSE SAS 语句 >;

当我们在创建 SAS 数据集时使用 IF 语句的格式之一，可以根据被处理的观测是否使 IF 条件表达式为真，来决定是否执行 THEN 后面的 SAS 语句。如果条件表达式为假，执行 ELSE 后面的语句，没有 ELSE 语句则执行 IF 语句的下一条语句。且 IF THEN/ELSE 语句还可多层嵌套，不过请注意此种语句的程序编写格式和匹配。

以下通过一个综合例子来说明以上语句的具体使用。

例 1-2　根据表 1-2 抽查的某班 8 个学生的体检结果新建数据集 chap1.example1_2，并实现如下操作：

（1）根据身高和体重计算每位学生的体重指数（BMI），已知体重指数等于体重（单位为 kg）除以身高（单位为 m）的平方。

（2）仅保留学生姓名、BMI 和性别。

（3）仅保留体重指数处在正常范围的学生信息（18<BMI<25）。

表 1-2　某班学生体检抽查结果

学　号（ID）	姓　名（Name）	性　别（Sex）	体　重（Weight）	身　高（Height）
01	姚籽萱	女	50.5	1.63
02	徐若黛	女	51	1.53
03	张　林	男	60	1.72
04	谢欣然	女	62	1.70

续表

学　号（ID）	姓　名（Name）	性　别（Sex）	体　重（Weight）	身　高（Height）
05	夏　天	女	54	1.67
06	刘子然	男	70	1.80
07	赵　赵	男	65	1.75
08	章　峰	男	84	1.68

注：体重的单位为 kg，身高的单位为 m。

编写如下程序（其在光盘中保存路径为“proc\chap1\example1_2”）：

```
data chap1.example1_2;
input ID $1-2 name $3-10 sex$11-12   +1 weight +1 height;
cards;
01 姚籽萱 女 50.5   1.63
02 徐若黛 女 51   1.53
03 张　林 男 60   1.72
04 谢欣然 女   62   1.70
05 夏　天 女 54   1.67
06 刘子然 男   70   1.80
07 赵　赵 男 65   1.75
08 章　峰 男 84   1.68
;
run;                          /*以上程序新建了包含了表格的数据集 chap2.example1_2*/
/*解答问题 A*/
data chap1.example1_2;
set chap1.example1_2;
BMI=weight/(height**2);      /*根据公式计算 BMI*/
run;
/*解答问题 B*/
data chap1.example1_2;
set chap1.example1_2;
keep ID sex BMI;             /*此行也可改写为 drop name weight height;*/
run;
/*解答问题 C*/
data chap1.example1_2;
set chap1.example1_2;
if BMI>18 && BMI<25;         /*保留 BMI 在 18～25 的观测*/
run;
```

1.3.3　对数据集排序

数据集的排序可应用 SORT 过程编程实现，该过程的语法结构如下：

```
PROC SORT <选项>;
BY <DESCENDING>  变量名;
RUN;
```

注意：DESCENDING 选项只对随后的一个变量起作用。如果省略 DESCENDING 指令，系统将默认指定升序排列。

例 1-3　将数据集 chap1.example1_2 对学生按其 BMI 的值降序排列，并另存为数据集 chap1.example1_3。

编写如下程序（其在光盘中的存储路径为“proc\chap1\example1_3”）：

```
proc sort data=chap1.example1_2 out=chap1.example1_3;
/*对数据集 chap1.example1_2 排序，将结果另存为 chap1.example1_3*/
by descending BMI;
 /*按照 BMI 的值降序排列，若需要升序排列则将 descending 改写成 ascending*/
run;
proc print data=chap1.example1_3;
run;
```

选择 Run|Submit 命令提交程序，打印输出结果如下：

The SAS System

Obs	ID	sex	BMI
1	02	女	21.7865
2	06	男	21.6049
3	04	女	21.4533
4	07	男	21.2245
5	03	男	20.2812
6	05	女	19.3625
7	01	女	19.0071

图 1-8　排序后输出结果

1.3.4　数据集纵向连接

数据的纵向连接指几个数据集中的数据纵向相加连接为一个新的数据集，它的记录数将是原来几个数据集中记录数的总和。此操作在 DATA 步中用 SET 指令实现。

Set 命令的语法格式如下：

```
set  数据集名称 1  数据集名称 2…数据集名称 n;
```

例 1-4　若已存在数据集 chap1.A、chap1.B 和 chap1.C（内容如表 1-3、表 1-4 和表 1-5 所示，存储在光盘中的路径分别为“data\chap1\A”、“data\chap1\B”、“data\chap1\C”）。

（1）将数据集 chap1.A 和 chap1.B 纵向连接成数据集 chap1.AB。

（2）将数据集 chap1.A 和 chap1.C 纵向连接成数据集 chap1.AC。

表 1-3　数据集 A

name	statistics	chinese
LR	89	92
HW	78	89
YJ	85	76
SJ	91	87

表 1-4　数据集 B

name	statistics	chinese
SL	78	87
LY	90	69

表 1-5　数据集 C

	name	statistics	English
1	YH	78	80
2	LJ	89	83

编写如下程序（其在光盘中的存储路径为“proc\chap1\set”）：

```
data chap1.AB;
set chap1.A    chap1.B;        /*合并数据集 chap1.A 和 chap1.B*/
run;

data chap1.AC;
set chap1.A chap1.C;           /*合并数据集 chap1.A 和 chap1.C*/
run;
```

选择 Run|Submit 命令提交程序，则新建数据集 chap1.AB 和 chap1.AC 如表 1-6 和表 1-7 所示。

表 1-6　数据集 chap1.AB

	name	statistics	chinese
1	LR	89	92
2	HW	78	89
3	YJ	85	76
4	SJ	91	87
5	SL	78	87
6	LY	90	69

表 1-7　数据集 chap1.AC

	name	statistics	chinese	English
1	LR	89	92	—
2	HW	78	89	—
3	YJ	85	76	—
4	SJ	91	87	—
5	YH	78	—	80
6	LJ	89	—	83

观察可知，数据集中的同名变量会自动合并（表 1-6 所示），对于其中不同名的变量，SAS

会在缺少相应数值的位置显示缺失（如表 1-7 所示）。

1.3.5　数据集横向合并

数据集的横向合并是指通过使用 MERGE 语句把两个及两个以上数据集中的两条或两条以上的观测合并为新数据集中的一条观测。它主要分为一对一合并和匹配合并。

一对一合并指把一个数据集中的第 k 条观测同另一个数据集中的第 k 条观测合并。新生成的数据集中的观测总数为这些数据集中观测个数的最大值。如果相应的某个数据集已没有观测，则相应的变量值为默认值。若在几个数据集中有共同的变量，则 SAS 默认使用最后出现的数据集中的变量值。

匹配合并与一对一合并最主要的区别在于它按照相同的关键变量合并。合并前必须把每个数据集根据关键变量排序。

例 1-5　若已存在数据集 chap1.C 和数据集 chap1.D（内容如表 1-5 和表 1-8 所示），将它们按照关键变量 name 横向连接成数据集 chap1.CD。

表 1-8　数据集 chap1.D

name	chinese
YH	89
LJ	78
HL	92

编写如下程序（其存储在光盘中的路径为“proc\chap1\merge”）：

```
proc sort data=chap1.C out=chap1.C;
by name;
run;
proc sort data=chap1.D out=chap1.D;
by name;
run;
/*以上对数据集 chap1.C 和 chap1.D 按照关键变量 name 升序排列*/
data chap1.CD;
merge chap1.C    chap1.D;          /*横向合并数据集 chap1.C 和 chap1.D*/
by name;                           /*设置关键变量为 id*/
run;
```

选择 Run|Submit 命令提交程序，则新建数据集 chap1.CD 如表 1-9 所示。

表 1-9　数据集 chap1.CD

	name	statistics	English	chinese
1	HL	—	—	92
2	LJ	89	83	78
3	YH	78	80	89

1.3.6 数据集转置

数据集的转置即把 SAS 数据集的行列互换。可应用 SAS 的 TRANSPOSE 过程完成此功能，该过程一般使用如下格式：

```
PROC TRANSPOSE <<DATA=输入数据集  OUT=转置数据集><选项列表>;
VAR  变量列表;
ID  变量;
COPY  变量列表;
RUN;
```

TRANSPOSE 过程从读取的数据集中创建主要包含以下三类变量的新数据集：

- 由原数据集中的观测转置后创建的新变量，即转置变量，如_NEME_、COL1、COL2、COL3…COLn。

说明：用户可自设转置变量名的方法有：

（1）通过 PROC TRANSPOSE 语句的选项 OLDNAME=NEWNAME 将原数据集中的变量名改为新变量 NEWNAME。

（2）通过选项 prefix=NO，修改默认的新变量名 COL1、COL2、COL3 为 NO1、NO2、NO3。

（3）用 ID 命令定义原数据集中某一变量的对应取值来代替新变量名称 COL1、COL2、COL3。

- 从原数据集中复制过来的变量，使用 COPY 语句定义这个变量，COPY 的变量与原数据集中的变量有相同的变量名和值。
- 为识别新数据集中每条观测的来源，用 ID 语句定义的新变量。

例 1-6 已知数据集 chap1.A（内容如表 1-3 所示，其存储在光盘中的路径为“data\chap1\A”），将其转置为新数据集 chap1.TA，并且将原数据集中的变量 name 变为 course、列变量改为学生姓名。

编写如下程序（其存储在光盘中的路径为“proc\chap1\example1_6”）：

```
proc transpose data=chap1.A out=chap1.TA name=course;
/*将数据集 chap1.A 转置为新数据集 chap1.TA，变量名 name 改为 course*/
var statistics chinese;                /*指定转置变量*/
id name;
run;
```

选择 Run|Submit 命令提交程序，新建转置后的数据集 chap1.TA 如表 1-10 所示。

表 1-10 数据集 chap1.TA

	name	statistics	English	chinese
1	HL	—	—	92
2	LJ	89	83	78
3	YH	78	80	89

1.3.7 数据集的导入与导出

SAS 系统的 Import Wizard 和 Export Wizard 提供了便捷的菜单操作方式实现数据集的导入与导出。以下通过例 1-7 详述使用方法。

例 1-7　已知 SPSS 数据文件 height.sav（其在光盘中的存储路径为 chap1/data/height.sav）。

（1）将其导入 SAS 系统，存储成数据集 chap1.example1_7。

（2）将数据集 chap1.example1_7 导出成 Excel 表格 height.xls。

问题（1）的操作过程如下。

步骤一：选择菜单 File|Import，弹出如图 1-9 所示对话框，Standard data source（标准数据）选项下的 Select a data source from the list 列出了能通过 Import Wizard 过程导入 SAS 系统的标准数据格式，本例中选择 SPSS File（*.sav），单击 Next 按钮进入如图 1-10 所示界面。

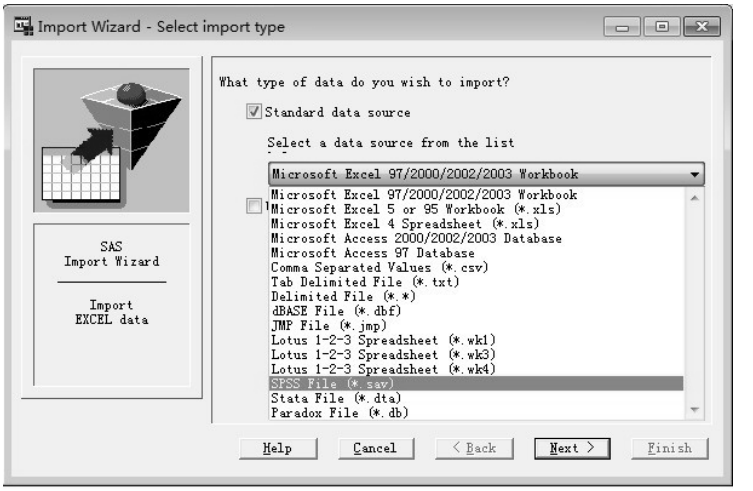


图 1-9　导入数据——选择数据类型

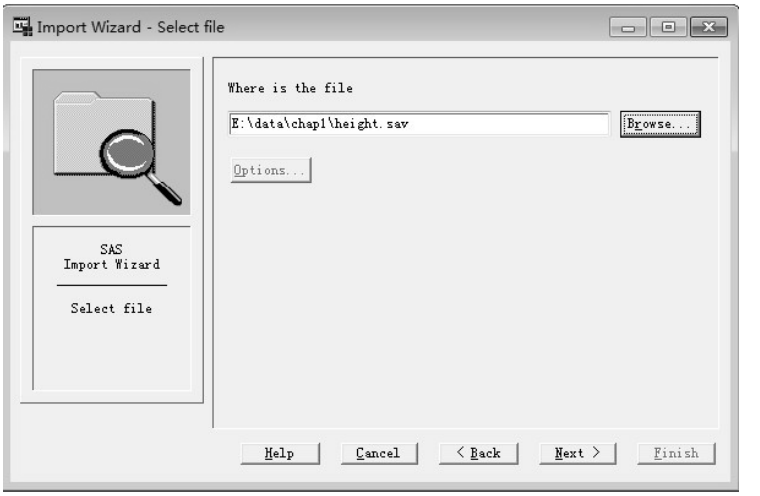


图 1-10　导入数据——指定数据文件

步骤二：在 Where is the file（文件路径）栏中输入需要导入文件在 PC 上的存储路径，或者单击 Browse（浏览）按钮在弹出的对话框中选择指定文件，单击 Next 按钮，进入图 1-11 所示界面。

步骤三：在 Choose the SAS destination（定义存储的 SAS 数据集）下的 Library（逻辑库）下拉列表中指定存储的逻辑库名 CHAP1，在 Member（对象）下拉列表中指定存储的数据集

EXAMPLE1_7（注意，SAS 数据集名不区分大小写，因此 CHAP1.EXAMPLE1_7 与 chap1.example1_7 指向同一个文件）。此时可以直接单击 Finish（完成）按钮实现数据集导入。

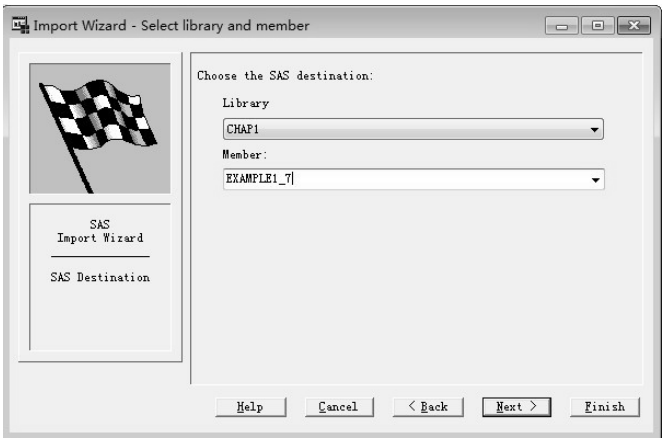


图 1-11　导入数据——确定数据集名

注意：若单击图 1-11 所示对话框上的 Next 按钮，在弹出的对话框中指定完成数据导入的 SAS 数据集的程序名及存储位置，此后可直接打开该程序（如以上操作的 SAS 程序如图 1-12 所示），更改 SAS 数据集名和原始数据位置即可快捷实现数据导入。此操作适用于需导入多个同一类别的数据文件。

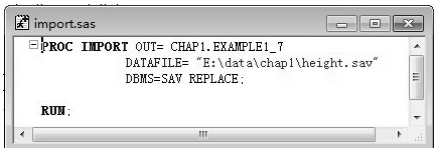


图 1-12　导入数据——SAS 程序

问题（2）的操作过程如下。

步骤一：选择菜单 File|Export，弹出如图 1-13 所示对话框，在 Library（逻辑库）下拉列表中选择 CHAP1，在 Member（对象）下拉列表中选择数据集 EXAMPLE1_7，单击 Next 按钮弹出如图 1-14 所示对话框。

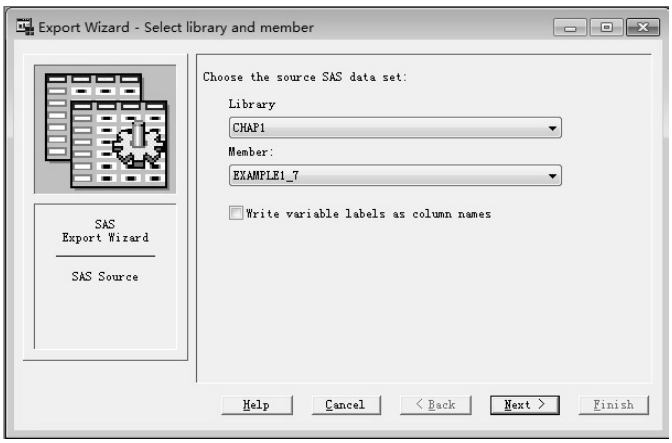


图 1-13　导出数据——选择数据集

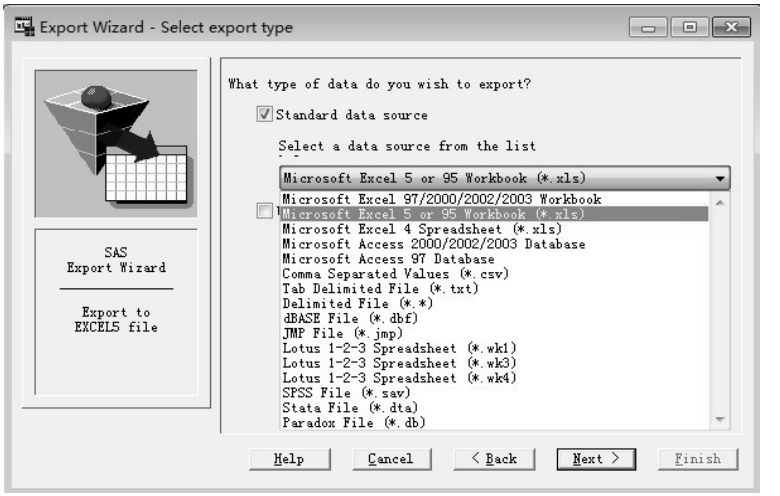


图 1-14　导出数据——选择数据类型

步骤二：在此选择导出数据格式 Microsoft Excel 5 or 95 Workbook(*.xls)（EXCEL 工作簿），单击 Next 按钮进入图 1-15 所示对话框。

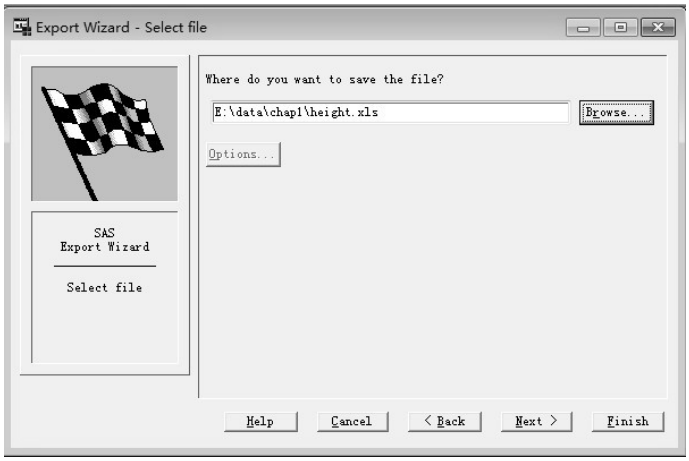


图 1-15　导出数据——确定存储路径

步骤三：确定导出文件存储路径和文件名“E:\data\chap1\height.xls”，单击 Finish（完成）按钮完成数据导出。

注意：与导入文件类似，导出过程可保存完成操作过程的 SAS 程序，更改相应部分实现数据的快速导出。

1.4　SAS 帮助系统介绍

SAS 帮助文档提供了非常详细的 SAS 系统的帮助信息，便于用户在使用过程中随时查询相关信息。以下简介 SAS 帮助文档的使用。

通过以下途径可以打开 SAS 帮助系统，如图 1-16 所示。

- 选择菜单栏中 Help|SAS Help and Documentation 命令。
- 单击工具栏上的按钮。
- 在命令栏中输入“help”，然后按回车键。
- 按下 F1 功能键。

图 1-16　SAS 帮助系统主界面

SAS 帮助系统界面的左侧有 4 个选项卡，分别为“目录”、“索引”、“搜索”和“收藏夹”。在“目录”选项内以文件夹目录树的形式依次列出了 SAS9.2 系统和以往版本相比更新的部分（What’s New in SAS9.2）、快速学习 SAS 系统的教程（Learning to Use SAS）、在不同的系统（UNIX、Windows、openVMS on HP Integrity Servers、z/OS）运行 SAS 的向导（Using SAS Software in Your Operating Environment）及 SAS 的产品介绍（SAS Products）。

在 SAS Products 目录下可以查询各个统计过程的具体介绍，如需要了解 SAS/STAT 模块下的 ANOVA 过程的具体介绍，双击 SAS/STAT 文件夹，选择子目录下的 SAS/STAT User’s Guide（SAS/STAT 模块用户向导），再在其子目录下选择 The ANOVA Procedure，则 SAS 帮助系统的右界面将显示出 ANOVA 过程的详细介绍（如图 1-17 所示），主要内容有简介过程（Overview）、进入过程（Getting Started）、语法介绍（Syntax）、细节（Details）、例题（Examples）和参考文献（References）。此时用户既可单击左侧目录树下的标题进入相应介绍，也可以直接单击右侧主界面上列出的详细标题查询相关信息。

以上直接根据目录查询是基于用户已知统计过程所属的模块，若用户实现没有这一信息，可以尝试使用“搜索”功能：单击“搜索”选项卡（如图 1-18 所示），在空白栏输入关键字，再

单击“列出主题”按钮，则将显示所有匹配的主题，单击相应的主题即可查看内容。

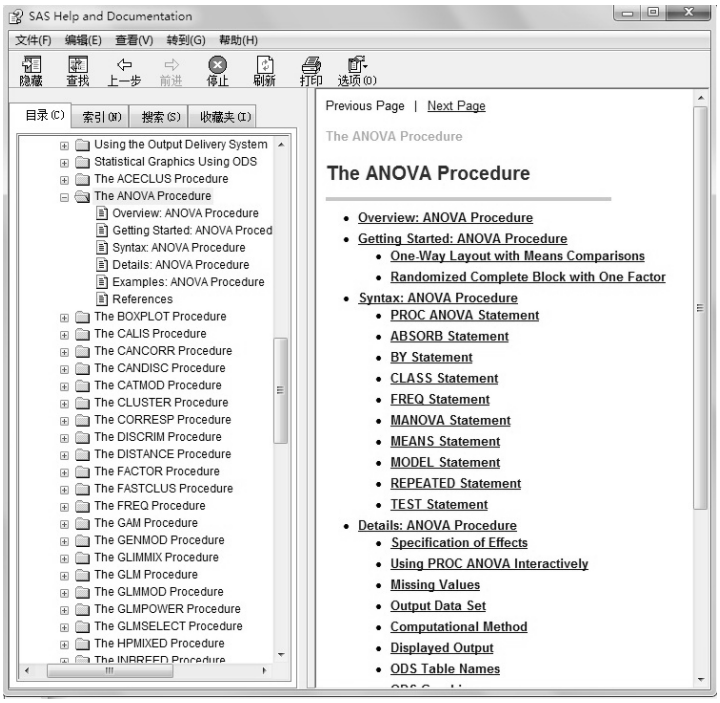


图 1-17　查询过程示意

若用户希望保存相应的查询内容以便于下次查看，可以单击“收藏夹”选项卡，再单击“添加”按钮将当前主题添加到收藏夹下。

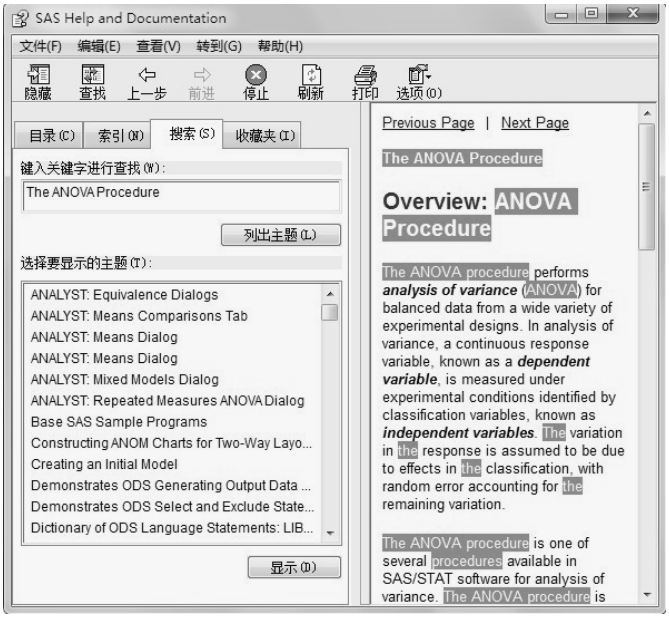


图 1-18　搜索主题

在实际的工作和学习中，遇到的 SAS 难题，除了求助于帮助系统，还可以考虑和更多的 SAS

用户交流，在此推荐一些 SAS 的在线帮助系统、交流论坛和优秀 SAS 人的博客。

- SAS 在线帮助系统：http://support.sas.com/index.html，在此可以查阅到 SAS 软件的最新动态、SAS 公司推出的新书推介等。
- 人大经济论坛 SAS 子版块：http://bbs.pinggu.org/forum-68-1.html，可以在此交流 SAS 难题、共享 SAS 学习心得和资源。
- COS 统计之都 SAS 子板块：http://cos.name/cn/forum/14。
- SAS 中文论坛：http://www.mysas.net/forum/viewforum.php?f=8。
- SAS 博客列表：http://saslist.com/，优秀 SAS 人的博客圈子。

第 2 章　SAS 编程简介

SAS 程序设计语言是一种专用的数据处理、统计计算语言，它不仅包含一般高级语言编程能力，而且扩充了许多数学、统计学方面的函数，且具备较强的数据管理功能。应用 SAS 软件以编程的方式进行数据整理和统计分析，过程透明、修改便捷，尤其适用于菜单操作不能完成的复杂统计分析。

本章首先简单介绍 SAS 程序的构成、基本语法规定，并介绍 SAS 数据步和过程步中常用的语句的使用意义，再简要介绍 SAS 系统自带的函数，最后概述交互式（ODS）输出系统，以及简介 SAS 宏编程基础。

2.1　SAS 程序简介

2.1.1　SAS 程序构成

SAS 统计分析程序主要包括两大步骤。

数据步：将用于分析的外部数据整理成 SAS 数据集。数据步由关键字 DATA 引出。

过程步：对 SAS 数据集进行调用、进行各类数据统计分析。过程步由关键字 PROC 引出。

注意：直接由数据步生成数据集适用于数据较少的情形，否则用户可考虑直接将外部数据导入成 SAS 数据集（详见 1.3.7 节）。

2.1.2　SAS 程序基本规定

以下为书写 SAS 程序的一些基本规定：

- SAS 程序以西文状态下的“;”作为结束符（注意：不能使用中文分号“；”）。
- SAS 程序命令中一般不区分大小写字母（注意：仅在作为数据的字符串中区分大小写）。
- 数据步和过程步各自包含若干条语句，多条语句可写在一行，但建议每条语句单独分行从而使程序具备较好的可读性。

2.2　数据步中基本语言介绍

SAS 的 DATA 步语法格式如下：

```
DATA 数据集名;
INPUT 变量名 1[$] 变量名 2[$] 变量名 3[$];    /*若为字符型变量后面加符号$*/
其他数据步语句;
CARDS;          /*注意，当变量中包含有“;”符号时，此处用 CARDS4*/
源数据行
;               /*注意，此处的分号要另起一行，若以上为 CARDS4 时，此处应该为“;;;;”*/
RUN;
```

以下将详细介绍 DATA 步中常用的 INPUT 数据输入语句、赋值语句、循环结构及分支结果。

2.2.1 INPUT 语句

DATA 步中的 INPUT 语句主要用于确定 SAS 数据集中包含的变量类型、排列次序及实现数据的导入。INPUT 语句读入 CARDS 语句下面的数据，而对于已存在的永久 SAS 数据集，可以用 SET、MERGE、UPDATE 等命令来实现数据集的调用、合并或更新。

INPUT 语句使用格式如下。

1．自由格式

此为最简单的数据输入格式，即按顺序列出每个观测的变量名，中间用空格隔开。注意变量若为字符型格式，需要在后面紧跟一个取地址符号$。$符号和变量名之间可以相连或用空格隔开。一般使用格式如下：

```
INPUT 变量名 1[$] 变量名 2[$] 变量名 3[$]…变量名 n[$];
```

例如以下程序：

```
data record;
input name$ height weight;
/*输入字符型变量 name、数值型变量 height、weight，变量间用单个空格隔开*/
cards;
叶子 165 53
木头 176 63
;
Run;
```

自由格式输入使用简单，输入时数据并不严格要求上下对齐，只需知道变量的次序，而不必确定它的具体列数，但是它有以下限制条件：

- 数据项之间至少需要用一个空格分隔。
- 字符型数据中间不能有空格。
- 缺失数据必须用小数点表示。
- 在 INPUT 语句中必须列出所有的变量名。

2．列标识方式

注意在 INPUT 关键字后面列出变量名外，还要在它后面列出该变量在数据行所占据的起始位置和结束位置。格式如下：

```
INPUT 变量名 1[$] 起始列-终止列 变量名 2[$] 起始列-终止列…变量名 N[$] 起始列-终止列
```

例如，可将程序 class 改写成程序 class1：

```
data record;
input name$ 1-4    height 6-8 weight 10-12;
/*指定字符型变量 name 占 1～4 列、数值型变量 height、weight 分别占 6～8 和 10～12 列*/
cards;
叶子  165 53
木头  176 63
;
Run;
```

注意到若各数据行的各个数据项是上下对齐的，可采用此种输入方式，其输入条件如下：

- 要求数据行各项上下对齐。
- 字符型数据中可以嵌套空格。
- 缺失值可用空格补齐。
- 可以只输入数据行中某些项。

3．格式输入

格式输入用于输入有特殊格式的数据，如日期、带小数的数字、含有空格的字符串等。

常用的 SAS 变量输入格式描述符说明如下。

W.：宽度为 W 位的标准数字。例如，8.代表数值型数据长为 8 个字符，且无小数。

W．D：含有小数点的标准数字，且它的总长度为 W 位。例如，8.2，效果为-5623.52。

注意：小数点、负号各占一位。

$W.：长度为 W 的标准字符串。例如，$10.指字符串的长度为 10 位。

COMMAW.D：长度为 W 的数字，与 W.D 的区别在于每三位整数部分用一个逗号隔开，且逗号占一位。例如，COMMA11.1，效果为 1,487,935.2。

$CHARW：宽度为 W，且含有空格的字符串。

常用的日期输入格式如下：

MMDDYY6.：月日年 6 位，如 012787。

MMDDYY8.：月日年 8 位，如 01/27/87 或 01-27-87 或 01271987。

DDMMYY6.：日月年 6 位，如 270187。

YYMMDD6.：年月日 6 位，如 870127。

DATE7.：日月年 7 位，如 27JAN87。

DATE9.：日月年 9 位，如 27JAN1987。

MMDDYY10.：月日年 10 位，如 01/27/1987 或 01-27-1987。

2.2.2　赋值语句

在 SAS 中用赋值语句计算表达式的结果，并将结果保存在等式左侧的变量中。SAS 数据步的程序中的计算是用表达式完成的。赋值语句的一般使用格式如下：

```
变量名=表达式;
```

例如：

```
BMI=height/(weight)^2
Z=Log(Y)**2
Isfact=(thief="T");
```

在以上的表达式中，变量 height、weight、Y、thief 必须存在于对应的数据集中，运算结果分别被保存在变量 BMI、Z、Isfact 中。

以下分别介绍 SAS 中的主要常量、变量、运算符。

1．SAS 系统中的常量

SAS 系统中的常量主要包括字符型、日期型、数值型和时间型数据，范例如表 2-1 所示。

表 2-1　SAS 系统中的常量

类　别	范　例	特征或用法注意
字符型	“优秀”、’English’	用单引号定界的字符常量和双引号定界的字符串
日期型	‘01/27/87’d	要求在表示日期的字符串后面加字母 d，中间不空格
数值型	90、−9.7、5.7E−4	包括整数、定点实数和科学计数法实数
时间型	‘14:25’t	要求在表示时间的字符串后面加字母 t
日期时间型	‘14JAN2007:18:45:51’dt	要求在表示日期时间的字符串后面加字母 dt

注：SAS 系统中的缺失值通常用一个单独的小数点表示。

2．SAS 系统中的变量

SAS 系统中的变量基本分为字符型和数值型。日期、时间等用数值型数据保存，若未定义字符型变量的长度，系统默认其占 8 字节。也可以用 LENGTH 语句指定变量长度，语句使用格式如下：

```
Length 变量名$长度;
```

例如：

```
Length country $30;      /*规定变量 country 占用 30 字节*/
```

3．SAS 系统中的主要运算符

算术运算符：主要包括+（加）、−（减）、*（乘）、/（除）、**（乘方）。在运算中*、/、**的优先级别高于+、−。若出现同级运算，则从左至右依次计算。

逻辑运算符：SAS 系统中的逻辑运算符主要有 AND、OR 和 NOT，如表 2-2 所示。

表 2-2　SAS 逻辑运算符

操 作 符	符　号	意　义	运 算 格 式	特　征
AND	&	与	运算对象 1 AND 运算对象 2	运算对象全为真，运算结果为真
OR	|	或	运算对象 1 OR 运算对象 2	一个运算对象为真，运算结果为真
NOT	^	非	NOT 运算对象	运算结果和运算对象本身真假性相反

关系运算符：SAS 中的关系运算符的具体内容如表 2-3 所示。

表 2-3　SAS 关系运算符

操 作 符	意　义	符　号
LT	小于（Less Than）	<
GT	大于（Greater Than）	>

续表

操　作　符	意　　义	符　　号
EQ	等于（Equal）	=
LE	小于等于（Less Equal）	<=
GE	大于等于（Greater Equal）	>=
NE	不等于（Not Equal）	^=
IN	等于列举中一个	

当运算对象满足运算符所指定的关系时，逻辑结果为真，其一般使用格式为：

```
运算对象 1 运算符 运算对象 2
```

2.2.3　循环语句

实际应用中，当需要建立符合特定需求的 SAS 数据集时，将用到 SAS 循环控制语句：DO 循环、DO WHILE 循环、DO UNTIL 循环。

1．DO 循环

语法格式：

```
DO 计数变量=起始值 TO 结束值 BY 步长;
循环体…;
END;
```

其中循环体由一个或多个语句构成，程序控制由计数变量被赋的起始值开始，循环体中语句每执行一次，则计数变量=计数变量+步长。重复执行循环体，直到计数变量超过指定的结束值。在循环体中可以用 LEAVE 语句来跳出循环，使用 CONTINUE 语句结束本轮循环，调整计数变量进行下一轮循环。

下面用例 2-1 演示以上语句的用法。

例 2-1　新建包含变量 x 和 y 的数据集 chap2.example2_1，x 取 5～30 的 5 的倍数，y 为 x 的自然对数值，且 y 的取值小于 3。

编写如下两段程序分别用 DO 循环结合 LEAVE 和 CONTINUE 语句完成分析（程序在光盘中的存储路径为“proc\chap2\example2_1”）：

```
/*方法一：DO 循环结合 LEAVE 语句*/
data chap2.example2_1;
do x=5 to 30 by 5;                  /*设置 x 的起始值为 5，终止值为 30，步长为 5*/
y=log(x);                           /*取 y 值为 x 的自然对数*/
if y>3 then leave;                  /*若 y 的值大于 3 则结束循环*/
output;
format y 8.5;                       /*定义 y 的输出形式为总长度为 8 位，小数部分占 5 位*/
end;                                /*结束循环*/

/*方法二：DO 循环结合 CONTINUE 语句*/
data chap2.example2_1b;
do x=5 to 30 by 5;                  /*设置 x 的起始值为 5，终止值为 30，步长为 5*/
y=log(x);                           /*取 y 值为 x 的自然对数*/
```

```
if y>3 then continue;          /*若 y 的值大于 3 则跳出循环*/
output;
format y 8.5;                  /*定义 y 的输出形式为总长度为 8 位，小数部分占 5 位*/
end;                           /*结束循环*/
```

选择 Run|Submit 命令提交以上任意一段程序，新建数据集 chap2.example2_1，如图 2-1 所示。

	x	y
1	5	1.60944
2	10	2.30259
3	15	2.70805
4	20	2.99573

图 2-1　数据集 chap2.example2_1

2．DO WHILE 循环

DO WHILE 循环语法格式如下：

```
DO WHILE 循环的语法结构;
 DO WHILE (循环继续条件);
    循环体语句;
END;
```

在语句开始执行时，程序通常首先判断循环条件表达式的逻辑结果是否为真，若为真则继续执行循环体语句，若为假则循环结束，循环体语句每执行一次，检验循环条件一次。下面用例 2-2 具体说明此语句的使用。

例 2-2　新建包含变量 x 和 y 的数据集 chap2.example2_2，y 的初始值为 100，x 取 2～10 之间的 2 的倍数，每次循环取 y 的值为前一次循环结束后 y 的值和当次循环 x 值的乘积，若 y 的值大于 1000 则结束循环。

编写如下程序（其在光盘中的存储路径为“proc\chap2\example2_2”）：

```
data chap2.example2_2;
y=100;                                  /*取 y 的初始值为 100*/
do x=2 to 10 by 2 while (y<3000);       /*当 y 的值小于 3000 时，继续循环*/
y=y*x;                                  /*y 的值为前一次循环结束后 y 的值和当次循环 x 值的乘积*/
output;
end;
run;
```

选择 Run|Submit 命令提交程序，新建 SAS 数据集 chap2.example2_2，如图 2-2 所示。

	y	x
1	200	2
2	800	4
3	4800	6

图 2-2　结果输出

注意：本程序进行了三次循环，当 y 的值为 4800，经检验不满足“y 的值小于 3000”这一循环继续条件后才结束循环。所以经 SAS 语句 y=y*x;计算的 y 最后一个值 4800 包含于数据集中。

3．DO UNTIL 循环

DO UNTIL 语句一般使用格式为：

```
DO UNTIL(循环退出条件);
循环体语句…;
END;
```

其中循环体语句是 SAS 语句，循环体退出条件为逻辑表达式。语句开始执行，首循环体每执行一次都要判断是否满足循环退出条件，若循环退出条件表达式的逻辑结果为真，则退出循环，否则重复执行循环体语句。

下面用例 2-3 具体说明此语句的用法。

例 2-3　改用 DO UNTIL 语句编程完成例 2-2 的问题。

编写如下程序（其在光盘中的保存路径为“proc\chap2\example2_3”）：

```
data chap2.example2_2;
y=100;                                  /*取 y 的初始值为 1*/
do x=2 to 10 by 2 until (y>=3000);      /*执行循环直到 y 的值大于等于 3000*/
y=y*x;                                  /*y 的值为前一次循环结束后 y 的值和当次循环 x 值的乘积*/
output;
end;
run;
```

选择 Run|Submit 命令提交程序，则新建与 SAS 数据集 chap2.example2_2 完全一样的数据集。

以上详述的 SAS 语言中常用的三种循环语句 DO、DO WHILE、DO UNTIL 都可以在 DATA 步中实现循环操作，它们的主要区别在于：DO 循环通常不设定限制条件；DO WHILE 和 DO UNTIL 循环都设置了循环条件；DO WHILE 的条件在循环体开头，而 DO UNTIL 的条件在循环体的结束。

2.2.4　分支结构

分支结构主要用于满足某条件，则执行某操作的情形，在 SAS 语言中，主要用 IF 语句和 SELECT 语句来实现此功能。1.3.2 节在介绍筛选数据集的观测时已初步介绍了 IF 语句的使用格式和用法。以下介绍 SELECT 语句。

SELECT 语句结构有两种用法，一般根据离散型变量分类采用用法一，而根据连续性变量分类采用用法二。

用法一的一般格式为：

```
SELECT(选择表达式);
        WHEN(值列表 1) SAS 语句 1;
          …
        WHEN(值列表 K) SAS 语句 K;
          …
        OTHERWISE  语句 N;
    END;
```

此种格式中的“选择表达式”是一个计算结果为数值或字符的表达式，“值列表”由一个或多个项构成（多项用逗号隔开）。系统首先计算选择表达式的结果，然后从上到下依次将选择表达式的结果与值列表中的各项相匹配，若匹配成功则进入对应的语句 K 并退出 SELECT 结构。若匹配不成功则执行 OTHERWISE 语句 N。例如以下程序：

```
select(Judge);                          /*定义分析变量 Judge*/
when(1,5)   Type="谷物类";              /*若变量 Judge 取值为 1 或 5，变量 Type 取值为“谷物类”*/
```

```
when(2,7)   Type="蔬果类";          /*若变量 Judge 取值为 2 或 7，变量 Type 取值为“蔬果类”*/
otherwise Type= "其他类";           /*若变量 Judge 取值不为以上数值（1、2、5 或 7），变量 Type
                                      取值为“其他类”*/
end;
```

用法二的一般格式为：

```
SELECT;
        WHEN(条件 1) SAS 语句 1;
                …
        WHEN(条件 K) SAS 语句 K;
                …
        OTHERWISE  语句 N;
    END;
```

此种格式在每一个 WHEN 语句后面都指定一个条件。程序执行首先进入第一个满足条件的 WHEN 所指定的语句 K。若所有条件都不满足则执行 OTHERWISE 语句 N。

下面用例 2-4 介绍说明 SELECT 语句的第二种用法。

例 2-4　已知根据表 1-2 建立数据集 chap2.example2_4，请计算每位体检者的体重指数（BMI）并且根据 BMI 的取值划分等级：BMI<18.5——“轻体重”；18.5≤BMI≤24——“健康体重”；24≤BMI<28——“超重”；BMI≥28——“肥胖”。

编写如下程序（其在光盘中的存储路径为“proc\chap2\example2_4”）：

```
Data chap2.example2_4;
set chap2.example2_4;              /*导入数据集*/
length type$ 12.;                  /*指定变量 type 的长度*/
BMI=weight/(height**2);            /*根据公式计算 BMI*/
select;                            /*选择语句*/
when(BMI<18.5) type='轻体重';
when(18.5<=BMI< 24) type='健康体重';
when(24 <=BMI< 28) type='超重';
when(BMI>=28) type='肥胖';
end;
run;
```

选择 Run|Submit 命令提交程序，得到数据集 chap2.example2_4，如图 2-3 所示。

	ID	name	sex	weight	height	type	BMI
1	01	姚籽萱	女	50.5	1.63	健康体重	19
2	02	徐若缇	女	51	1.53	健康体重	22
3	03	张　林	男	60	1.72	健康体重	20
4	04	谢欣然	女	62	1.7	健康体重	21
5	05	夏　天	女	54	1.67	健康体重	19
6	06	刘子然	男	70	1.8	健康体重	22
7	07	赵　赵	男	65	1.75	健康体重	21
8	08	章　峰	男	84	1.68	肥胖	30

图 2-3　数据集 chap2.example2_4

2.3　过程步中基本语句介绍

用户在数据步中编写 SAS 程序来读入、处理和描述数据以创建符合要求的 SAS 数据集，此

后基于数据集可调用 SAS 的过程步进行分析。对于许多常用的和标准的统计方法，可以调用 SAS 系统自带的相应过程步，只有非常特殊的统计方法才需要自主编程。

SAS 过程步的一般格式为：

```
PROC 过程名 [DATA=输入数据集] [选项];
过程语句 1[选项];
…
过程语句 N[选项];
RUN;
```

过程名指定 SAS 过程的名字，如对数值变量计算简单描述统计量的过程名 MEANS 等。选项规定在分析过程中的特定计算要求。不同的过程规定的选项也不一样，而以下三种选项在不同过程使用格式相同：

- Keyword
- Keyword=数值
- Keyword=数据集

Keyword 是关键字，第一种选项格式是某个具体过程进一步要求某个关键字；第二种选项格式是某个具体过程要求某个关键字的值，值可能是数值或字符串；第三种选项格式是某个具体过程要求输入或输出数据集。

本节概述在 SAS PROC 步中的 VAR、MODEL 等常用语句，在接下来的章节中将会在不同的统计过程中应用这些语句。

2.3.1　VAR、MODEL、BY、CLASS 语句

VAR 语句被用来指定需要分析的变量。若指定多个变量，变量间用空格隔开。其语句格式为：

```
VAR 变量列表;
```

例如：

```
VAR CHINSES ENGLISH MATH;
```

MODEL 语句用在统计建模中指定模型的形式。其语句格式为：

```
MODEL 因变量=自变量列表/选项;
```

例如：

```
MODEL Y=X1 X2;
```

BY 语句可用在对数据集进行分组处理，要求数据集事先按 BY 变量进行排序（用 SORT 语句等）。其语句格式为：

```
BY <DESCENDING> 变量 1 <…变量 K> ;
```

注意：BY 后面的变量排列的先后次序表示分组的先后次序。

CLASS 语句用来指定一些分类变量。其语句格式为：

```
CLASS 变量列表;
```

例如：

```
CLASS TYPE;
```

注意：CLASS 语句与 BY 语句的区别在于：使用 CLASS 语句不要求数据集事先按 CLASS 指定的变量排序，按指定变量的不同值进行分类计算和分析后，输出的分类结果列在一张报表

里。而 BY 语句在使用时，要求数据集事先按 BY 指定的变量排序，且输出的结果也按分组列出许多报表。

2.3.2 WHERE、FREQ、WEIGHT 语句

WHERE 语句用于选择输入数据集的子集进行分析。其语句格式为：

```
WHERE 逻辑表达式;
```

例如：

```
WHERE MATH<60 OR ENGLISH<60;        /*指定分析数学或英语成绩不及格的学生*/
```

FREQ 语句用来指定一个代表观测出现的频数变量。其语句格式为：

```
FREQ 变量列表;
```

例如：

```
FREQ FEMALE;
```

WEIGHT 语句用来指定一个代表观测权重的变量。其语句格式为：

```
WEIGHT 变量;
```

例如，在某些允许加权的分析中，其值和观测对应的方差成反比。

注意：如果在某个观测中，FREQ 变量的值小于 1，这个观测在分析中不使用；如果 FREQ 变量的值不是整数，仅取整数部分使用。注意 FREQ 语句和 WEIGHT 语句的区别。FREQ 变量表示观测出现的次数；WEIGHT 变量给出观测相应的权数。当每个观测的权数都是整数时，WEIGHT 语句也可用 FREQ 语句代替。

2.3.3 使用 OUTPUT、FORMAT、 LABEL、ID 语句

OUTPUT 语句经常用来指定输出结果存放的数据集。其语句格式为：

```
OUTPUT OUT=输出数据集名 关键字=变量名 关键字=变量名…;
```

用关键字定义输出变量名，等号为此关键字在输出数据集中的名称。

FORMAT 语句用于指定变量输出格式。其语句格式为：

```
FORMAT 变量名 1 格式描述 1 变量名 2 格式描述 2…;
```

例如：

```
FORMAT NAME$1-10 HEIGHT5.1 WEIGHT;
```

LABEL 语句用于为变量指定标签。其语句格式为：

```
LABEL 变量名='标签' 变量名='标签'…;
```

例如：

```
LABEL NAME='姓名' BIRTH='出生日期';
```

说明：在 DATA 步和 PROC 步都可以用到 FORMAT、LABEL 语句，但是在 DATA 步中规定的变量属性是永久的，在 PROC 步中规定的变量标签等只对本次过程有效。

ID 语句用来指定用于识别观测的一个或几个变量。使用了 ID 语句后，最左边的 SAS 默认输出的观测列（obs）被 ID 语句所指定的变量替代。

2.4　SAS 函数

SAS 系统中自带了许多数学和统计学方面的标准函数，这些函数可以直接应用到数据步的运算和赋值当中。

函数的调用方法非常简单，如计算 y 的自然对数，则表示成 log(y)即可。

本节仅介绍主要的 SAS 函数，其他函数可查看 SAS 帮助系统。（具体目录为：Base SAS Software| Using Base SAS Software|Working with the SAS Language|SAS Function| Function Categories）。读者还可参考《SAS 软件：Base SAS 软件使用手册》（高惠璇等编译，中国统计出版社出版）。以下为常见函数简介。

1．数学函数

在 SAS 系统中，数学函数主要包括三角函数、对数函数等，如表 2-4 所示。

表 2-4　SAS 系统中的主要数学函数

函 数 名	意　义
ABS(x)	求 x 的绝对值
MAX(x1,x2,…,xn)	求 x1 到 xn 中最大的一个
MIN(x1,x2,…,xn)	求 x1 到 xn 中最小的一个
MOD(x,y)	求 x 除以 y 的余数
SQRT(x)	求 x 的平方根
INT(x)	求 x 去掉小数部分后的结果
LOG(x)	求 x 的自然对数
LOG10(x)	求 x 的以 10 为底的对数
EXP(x)	指数函数
SIN(x)	求 x 的正弦
COS(x)	求 x 的余弦
TAN(x)	求 x 的正切
ARSIN(y)	计算函数 y=sin(x)的反函数，y 取[−1,1]
ARCOS(y)	计算函数 y=cos(x)的反函数，y 取[−1,1]
ATAN(y)	计算函数 y=tan(x)的反函数，y 取[−1,1]
SINH(x)	求 x 的双曲正弦
COSH(x)	求 x 的双曲余弦
TANH(x)	求 x 的双曲正切
ARSINH (x)	求 x 的反双曲正弦
ARCOSH（x）	求 x 的反双曲余弦
ARTANH(x)	求 x 的反双曲正切
GCD	返回一个或多个整数的最大公约数
LCM	返回能被一组数中的每个数整除的最小倍数

续表

函 数 名	意 义
ERF(x)	误差函数
SIGN	符号函数

2．字符函数

SAS 系统中常用的字符函数如表 2-5 所示。

表 2-5 SAS 系统中常用的字符函数

函 数 名	意 义
TRIM(s)	返回去掉字符串 s 的尾随空格的结果
UPCASE(s)	把字符串 s 中所有小写字母转换为大写字母后的结果
LOWCASE(s)	把字符串 s 中所有大写字母转换为小写字母后的结果
INDEX(s,s1)	查找 s1 在 s 中出现的位置。找不到时返回 0
RANK(s)	字符 s 的 ASCII 码值
BYTE(n)	第 n 个 ASCII 码值的对应字符
REPEAT(s,n)	字符表达式 s 重复 n 次
SUBSTR(s,p,n)	从字符串 s 中的第 p 个字符开始抽取 n 个字符长的子串
TRANWRD(s,s1,s2)	从字符串 s 中把所有字符串 s1 替换成字符串 s2 后的结果

3．日期和时间函数

在 SAS 系统中，日期数据以 1960 年 1 月 1 日为起始日，系统以距离起始日期的总天数记录实际日期。例如，1962 年 3 月 20 日被存储为 890，表示此日期与 1960 年 1 月 1 日相距 890 天。所以，当变量的值为日期或时间类型时，用户需指定变量的输入和输出格式。SAS 系统中常用的日期和时间函数如表 2-6 所示。

表 2-6 SAS 系统中常用的日期和时间函数

函 数 名	意 义
MDY(m,d,yr)	生成 yr 年 m 月 d 日的 SAS 日期值
YEAR(date)	由 SAS 日期值 date 得到年
MONTH(date)	由 SAS 日期值 date 得到月
DAY(date)	由 SAS 日期值 date 得到日
WEEKDAY(date)	由 SAS 日期值 date 得到星期几
QTR(date)	由 SAS 日期值 date 得到季度值
HMS(h,m,s)	由小时 h、分钟 m、秒 s 生成 SAS 时间值
DHMS(d,h,m,s)	由 SAS 日期值 d、小时 h、分钟 m、秒 s 生成 SAS 日期时间值
DATEPART(dt)	求 SAS 日期时间值 dt 的日期部分
INTFIT	返回两个日期之间的时间间隔
INTGET	返回基于三个日期值或日期时间值的时间间隔

4．分布密度函数、分布函数

作为统计计算语言，SAS 提供了多种概率分布的有关函数。分布密度、概率、累积分布函数等可以通过几种统一的格式调用，格式如表 2-7 所示。

表 2-7　求概率分布函数的值

求值的类型	格　　式
分布函数值	CDF('分布', x <, 参数表>)
密度值	PDF('分布', x <, 参数表>)
概率值	PMF('分布', x <, 参数表>)
对数密度值	LOGPDF('分布', x <, 参数表>)
对数概率值	LOGPMF('分布', x <, 参数表>)

其中，CDF 计算由'分布'指定分布的分布函数，PDF 计算分布密度函数值，PMF 计算离散分布的分布概率，LOGPDF 为 PDF 的自然对数，LOGPMF 为 PMF 的自然对数。函数在自变量 x 处计算。

分布类型取值可以为：BERNOULLI，BETA，BINOMIAL，CAUCHY，CHISQUARED，EXPONENTIAL，F，GAMMA，GEOMETRIC，HYPERGEOMETRIC，LAPLACE，LOGISTIC，LOGNORMAL，NEGBINOMIAL，NORMAL 或 GAUSSIAN，PARETO，POISSON，T，UNIFORM，WALD 或 IGAUSS 及 WEIBULL。在程序中调用这些函数名称时，可以只写前 4 个字母。除了用上述统一的格式调用外，SAS 还单独提供了常用的分布的密度、分布函数，如表 2-8 所示。

表 2-8　常用分布函数

形　　式	意　　义
PROBNORM(x)	标准正态分布函数
PROBT(x,df<,nc>)	自由度为 df 的 t 分布函数，可选参数 nc 为非中心参数
PROBCHI(x,df<,nc>)	自由度为 df 的卡方分布函数，可选参数 nc 为非中心参数
PROBF(x,ndf,ddf<,nc>)	F(ndf,ddf)分布的分布函数，可选参数 nc 为非中心参数
PROBBNML(p,n,m)	设随机变量 Y 服从二项分布 B(n,p)，此函数计算 P(Y,m)
POISSON((lambda,n)	参数为 lambda 的 Poisson 分布 Y n 的概率
PROBNEGB(p,n,m)	参数为(n,p)的负二项分布 Y m 的概率
PROBHYPR(N,K,n,x<,r>)	超几何分布的分布函数
PROBBETA(x,a,b)	参数为(a,b)的 Beta 分布的分布函数
PROBGAM(x,a)	参数为 a 的 Gamma 分布的分布函数
PROBMC	计算多组均值的多重比较检验的概率值和临界值
PROBBNRM(x,y,r)	标准二元正态分布的分布函数，r 为相关系数

5．分位数函数

分位数函数是概率分布函数的反函数。其自变量取值范围为 0～1。分位数函数计算的是分

布的左侧分位数。SAS 提供了 6 种常见连续型分布的分位数函数，如表 2-9 所示。

表 2-9　常见分位数函数

形　式	意　义
PROBIT(p)	标准正态分布左侧 p 分位数
TINV(p, df <,nc>)	自由度为 df 的 t 分布的左侧 p 分位数，可选参数 nc 为非中心参数
CINV(p,df<,nc>)	自由度为 df 的卡方分布的左侧 p 分位数，可选参数 nc 为非中心参数
FINV(p,ndf,ddf<,nc>)	F(ndf,ddf)分布的左侧 p 分位数，可选参数 nc 为非中心参数
GAMINV(p,a)	参数为 a 的 Gamma 分布的左侧 p 分位数
BETAINV(p,a,b)	参数为（a,b）的 Beta 分布的左侧 p 分位数

6．随机数函数

SAS 可以用来进行随机模拟，它提供了以下常见分布的伪随机数生成函数，如表 2-10 所示。

表 2-10　常用随机数函数

分 布 类 型	函 数 形 式	使 用 条 件
均匀分布	UNIFORM(seed)	seed 必须是常数，为 0，或 5 位、6 位、7 位的奇数
	RANUNI(seed)	seed 为小于 $2^{31}-1$ 的任意常数
正态分布	NORMAL(seed)	seed 为 0，或 5 位、6 位、7 位的奇数
	RANNOR(seed)	seed 为任意数值常数
指数分布	RANEXP(seed)	seed 为任意数值，产生参数为 1 的指数分布的随机数
二项分布	RANBIN(seed,n,p)	seed 为任意数值，产生参数为（n,p）的二项分布随机数
泊松分布	RANPOI(seed,lambda)	seed 为任意数值，产生参数为 lambda>0 的泊松分布随机数

7．样本统计函数

样本统计函数把输入的自变量作为一组样本，计算样本统计量。其调用格式为“函数名（自变量 1，自变量 2，…，自变量 n）”或“函数名（OF 变量名列表）”。例如，SUM 是求和函数，如果要求 x1、x2、x3 的和，可以用 SUM(x1,x2,x3)，也可以用 SUM(OF x1-x3)。这些样本统计函数只对自变量中的非缺失值进行计算。常用的样本统计函数如表 2-11 所示。

表 2-11　常用的样本统计函数

表 达 形 式	意　义	表 达 形 式	意　义
MEAN	均值	NMISS	缺失值的个数
MAX	最大值	SUM	求和
MIN	最小值	VAR	方差
N	非缺失数据的个数	STD	标准差
STDERR	均值估计的标准误差	CV	变异系数
USS	平方和	CSS	离差平方和
SKEWNESS	偏度	KURTOSIS	峰度

注意：数据集的存储一般是每行为一个个体的观测值，每列是个体的一个属性（变量），所以统计一般应该对列进行，而不是像这里对行进行，把各变量作为一个样本的各个观测处理。这里提供的函数主要用于进行一些自编程的计算。

2.5　ODS 输出系统

传统的 SAS 系统可以支持结果以 SAS 数据集（DATASETS）、SAS 索引文件（THE SAS LOGS）、简单汇总报表（A REPORT OR SIMPLE LISTING）、目录树（CATALOGS）文件、保存于其他数据集的扩展文件的形式输出。SAS9.2 版中新增和增强了“输出交付系统”（以下简称 ODS）的功能，它提供各种各样的格式化选择和输出目标。ODS 能够让用户几乎不受限制地选择以各种美观的格式报告和显示分析结果，使用 DOCUMENT 过程，可以生成多个 ODS 输出及控制文档内容的去留及显示格式，而不必反复运行过程或数据查询。ODS 支持输出的主要格式如下。

- RTF：Rich Text Format 多文本文件格式。
- OUTPUT：SAS 数据集格式。
- LISTING：传统的 SAS 汇总输出。
- HTML：超文本链接，网页格式。
- PRINTER：高分辨率打印输出格式。

ODS 输出 RTF 文档语法格式如下：

```
ODS LISTING CLOSE;
ODS RTF FILE='文件名.RTF' ;
…;
ODS RTF CLOSE;
ODS CLOSE;
```

ODS 输出系统将输出结果整理成 SAS 数据集语法格式如下：

```
ODS LISTING CLOSE;
ODS OUTPUT "输出文件名"="SAS 数据集";
…;
ODS OUTPUT CLOSE;
ODS CLOSE;
```

注意：如果用户不需要创建列表输出，在程序的开头写入指令（ODS listing close;）。因为 ODS 语言是全程语言，在 SAS 语句执行结束后，建议用户在程序的最后加上输出列表文件的指令（ODS listing;）。

以下举例说明使用 ODS 语句来控制 SAS 过程输出 RTF 格式。

例 2-5　已知要求将数据集 chap2.example2_5.sas7bdat 打印输出保存为 RTF 格式文档 example2_5.rtf，且在计算机中的存储路径为“E:\proc\chap2”。

编写如下程序（其在光盘中的保存路径为“proc\chap2\example2_5.sas”）：

```
ODS listing close;                                  /*关闭 SAS 列表输出*/
ODS rtf file=' E:\proc\chap2\example2_5.rtf';   /*定义输出文档名为 example2_5 及存储路径*/
proc print data=chap2.example2_4;
run;
```

```
ODS rtf close;                          /*结束将 SAS 运算结果以文本文档形式输出的过程*/
ODS listing;                            /*打开 SAS 列表输出*/
```

选择 Run|Submit 命令提交程序，在 Output 输出记录窗口将显示空白，在指定的路径下发现 RTF 文本文档，打开文件如图 2-4 所示。

Obs	ID	name	sex	weight	height	type	BMI
1	01	姚籽萱	女	50.5	1.63	健康体重	19.0071
2	02	徐若蕾	女	51.0	1.53	健康体重	21.7865
3	03	张　林	男	60.0	1.72	健康体重	20.2812
4	04	谢欣然	女	62.0	1.70	健康体重	21.4533
5	05	夏　天	女	54.0	1.67	健康体重	19.3625
6	06	刘子然	男	70.0	1.80	健康体重	21.6049
7	07	赵　赵	男	65.0	1.75	健康体重	21.2245
8	08	章　峰	男	84.0	1.68	肥胖	29.7619

图 2-4　RTF 文档中的结果

2.6　SAS 宏简介

SAS/BASE 提供了宏（MACRO）功能，通过创建宏变量和宏能够方便地完成替代重复文本，避免繁复分析操作；获取系统信息；产生程序执行过程中的参数；用宏变量在数据步和过程步间传递数据；产生与数据无关的程序，使程序具有可移植性。

本节简单介绍 SAS 的宏变量、宏函数，定义与调用宏。

2.6.1　SAS 宏变量

宏变量属于 SAS 宏语言，是可在 SAS 程序的除数据行以外的任何地方定义和使用的变量，它独立于数据集。宏变量的值保持不变，直到被明确改变。宏变量分成两类：一是用户定义的宏变量，二是 SAS 系统自带的自动宏变量。

一般使用宏语句%LET 定义宏变量，它的一般形式如下：

```
%LET  宏变量名=值;
```

宏变量的命名遵从 SAS 命名规则，它的值可以是固定的字符串、其他宏变量的引用、宏函数和宏调用。

引用一个宏变量的值的格式为：&宏变量名。

以下通过例 2-6 简单示范宏变量的使用。

例 2-6　若需要分析多个数据集中的同样的变量，并且打印出结果，为避免重复输入数据集名，编写程序 example2_6.sas 如下所示，其在光盘中的存储路径为“proc\chap2\example2_6.sas”。

```
%Let    data=example;
/*注意，example 代表指定分析的数据集，若打印不同的数据集，更改此处即可*/
Proc print    data=&data;                    /*第一次引用宏*/
Var       name height weight;
Title    "Display of Data Set &data";       /*第二次引用宏*/
Run ;
```

注意：读者已经注意到，在此处定义宏变量的最主要功能是文本替换，避免重复输入相同的文本。

SAS 的自动宏变量在系统启动时被创建，可以在 SAS 程序任何地方被引用，且在 SAS 系统退出前一直保持有效，表 2-12 为常用的自动宏变量。

表 2-12　常用的自动宏变量

自动宏变量	作　　用	范　　例
SYSDATE	本次 SAS 启动的日期	如 19AGU09
SYSDAY	本次 SAS 启动的星期	如 Thursday
SYSTIME	本次 SAS 启动的时间	如 15:41
SYSLAST	最新创建的数据集名字	如 chap2.example2_4
SYSDSN	最新创建的数据集两部分名字	如 chap2　　example2_4
SYSVER	使用 SAS 软件的版本	如 9.2
SYSSCP	返回用户主机系统的缩写	如 Liurong

2.6.2　创建和调用宏

SAS 的宏函数简称宏，它可以通过控制和循环语句来控制文本的输出，用户还可以定义宏参数实现宏的多次调用。宏的一般定义格式如下：

```
%MACRO 宏名称(宏参数);
宏实体;
%MEND 宏名称;
```

%MACRO 语句标志着宏的开始，紧接的是用户自定义的宏名称，根据实际需求可以定义宏参数，将宏执行过程中需要调用的宏变量传递进去。宏实体可为任意一个文本，SAS 语句或 SAS 步及宏变量、函数和这些实体的组合。%MEND 标志宏的结束。

SAS 中调用宏的形式如下：

```
%宏名称(参数)
```

例 2-7　现在需要对不同 SAS 数据集中的数值型变量进行描述性统计分析，为避免重复输入文本，编写程序如下所示，其在光盘中的存储路径为“proc\chap2\example2_7.sas”。

```
%MACRO  calldataset(procname=,dataset=,varname=);
ods listing off;
ods rft file='Analysis &varname of &dataset';
Proc   &procname   Data=&dataset ;
var &varname;
Run ;
ods rft close;
ods listing;
%MEND calldataset;
```

可以采用类似以下的语句调用宏 calldataset：

```
%MACRO(procname=means,dataset=A,varname=height);
%MACRO(procname=univariate,dataset=B,varname=number);
```

在执行第一次调用时，该程序将被宏处理器替换成以下程序：

```
ods listing off;
ods rtf file='Analysis height of A';
Proc means Data=A ;
var height;
Run ;
ods rtf close;
ods listing;
```

在%MACRO 和%MEND 语句中可以使用条件语句%IF—%THEN/%ELSE 语句和循环语句%DO—%END、%DO %UNTIL—%END、%DO %WHILE—%END 语句来产生更复杂的宏。虽然表达形式有点差异，但是语句的意义与其在 DATA 步中的使用类似。

SAS 宏的内容非常丰富，使用也非常灵活，在此的概述仅为抛砖引玉，读者在实际学习和工作中遇到更多实际问题时可以参考 SAS 帮助文档，笔者推荐阅读 SAS 公司出版的《SAS Programing Made Easy,second edition》，Michele M Burlew 著。

练习题

习题 2-1 已知变量 X 和 Y 之间的函数关系式为：$Y=0.5X^2+3X+5$（$X \in Z$，且 $X \leqslant 50$），请在以下条件下，新建包含 X 和 Y 的全部取值数据集 chap2.exercise2_1。

（1）限定 Y 的最大值为 200，用 DO LEAVE 结构编程求解；

（2）限定 Y 的最大值为 200，用 DO CONTINUE 结构编程求解；

（3）限定若 Y 的值大于 200 则结束循环，用 DO WHILE 结构编程求解；

（4）限定若 Y 的值大于 200 则结束循环，用 DO UNTIL 结构编程求解。

（本习题的解答程序在光盘中的存储路径为“proc\chap2\exercise2_1”。）

习题 2-2 已知面试官给 10 个应聘者的表现从专业技能、言行举止和思维活跃度三个方面对他们评分，具体数据如表 2-13 所示，相应的 SAS 数据集在光盘中的存储路径为“data\chap2\exercise2_2”。

（1）请计算应聘者的加权得分，其中专业技能的权重为 0.4，言行举止的权重为 0.3，思维活跃度的权重为 0.3；

（2）请按照如下规则根据应聘者的加权得分评定其等级：加权得分大于 90 分——“A”；加权得分在 80～90 分——“B”级；加权得分在 70～80 分——“C”级；加权得分低于 70 分——“D”级；

（3）将应聘者的编号（ID）、平均得分和评定等级打印输出成 RTF 文档 exercise2_5.rtf。

表 2-13 应聘者得分情况

编号（ID）	专业技能得分（Score1）	言行举止得分（Score2）	思维活跃度得分（Score3）
P1	87	86	89
P2	74	87	85
P3	89	97	90
P4	89	86	87

续表

编号（ID）	专业技能得分（Score1）	言行举止得分（Score2）	思维活跃度得分（Score3）
P5	95	82	84
P6	78	81	79
P7	84	86	84
P8	83	83	81
P9	75	93	94
P10	89	92	92

（本习题的解答程序在光盘中的存储路径为“proc\chap2\exercise2_2”。）

习题 2-3　已知需要对不同数据集中不同的变量进行频数分析，并且打印出最终分析结果，请据此编写避免重复输入文本的两类 SAS 宏。

（本习题的解答程序在光盘中的存储路径为“proc\chap2\exercise2_3”。）

第 3 章　SAS 菜单操作

本书第 2 章简介了 SAS 编程的基本规则，重点介绍了数据步的相关语言，目的是让用户学会建立满足特定要求的 SAS 数据集，为各项统计分析做准备。虽然 SAS 的用户以专业人员为主，并大多选择灵活的编程分析，但为了适应用户群体的多元化需求，SAS 也提供了界面友好、使用便捷的菜单操作方式。本章主要介绍 SAS/ASSIST、SAS/INSIGHT 和 SAS/Analyst 模块，这些模块的具体应用在后续章节会多次提到。

3.1　SAS/ASSIST 视窗介绍

3.1.1　SAS/ASSIST 概述

选择菜单 Solutions|ASSIST，则 SAS 系统将搜索系统中所有可用的 SAS 数据集。搜索完毕则弹出如图 3-1 所示对话框，此为有关于菜单呈现方式及是否需要开启其他界面的选项，建议直接采用默认设置，单击 Continue 按钮。若不希望在下次启动 SAS/ASSIST 时再次出现此对话框，取消选定 Show this window at startup（启动时显示本界面），单击 Continue 按钮，出现 SAS/ASSIST 工作界面，如图 3-2 所示。

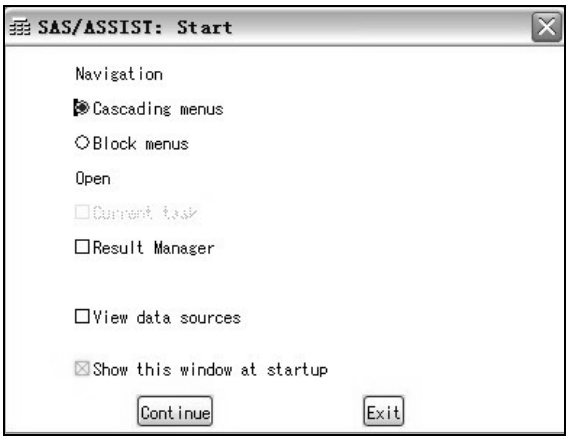


图 3-1　SAS/ASSIST 进入菜单

以下介绍 SAS/ASSIST 主界面的模块。

Data Mgmt：数据管理，包括创建、导入、导出、编辑、浏览、排序等关于数据集的操作及 SQL 语句查询功能。

Report Writing：报表制作。

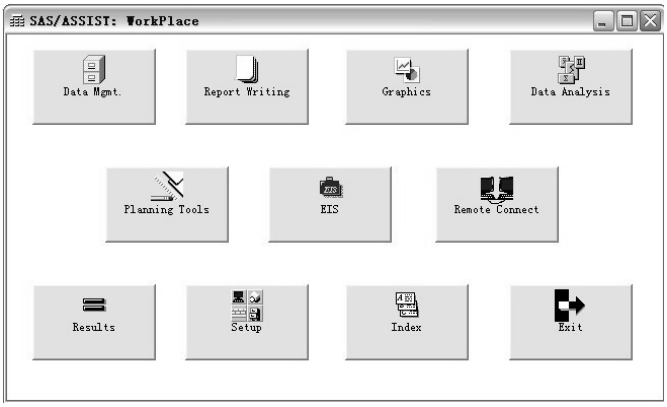


图 3-2　SAS/ASSIST 工作界面

Graphics：图形绘制，主要有条形图（Bar Charts）、饼图（Pie Chart）、各类散点图（Plots）、地图（Maps）等。

Data Analysis：数据分析，主要包括离散和连续性变量的描述性统计分析、回归分析、T 检验和方差分析、多元分析（包括主成分和因子分析）、质量控制、时间序列等。

Planning Tools：项目管理工具。

EIS：基于 SAS 的二次开发工具。

Remote Connect：管理本地计算机和远程计算机的 SAS 对话。

Results：管理存储在目录文件中的结果。

Setup：启动 SAS 文件管理、系统设置、帮助系统等。

Index：提供了 ASSIST 模块中所有功能的索引，适用于 SAS 用户入门 ASSIST。

Exit：退出 SAS。

虽然 SAS/ASSIST 提供了多个功能强大的模块，但最常用的是 Data Mgmt（数据管理）、Data Analysis（数据分析）和 Graphics（图形绘制）三个模块。下面简单介绍应用 ASSIST 的 Data Analysis 中 Summary Statistics（描述性统计分析）以菜单操作的方式完成分析。

3.1.2　SAS 实例——分性别描述某班学生英语成绩分布

例 3-1　已知某初中班级 50 名学生的英语成绩保存在数据集 chap3.example3_1 中（部分数据如图 3-3 所示，id 代表学号；sex 代表性别：1——男性，0——女性；score 代表英语成绩），数据集在光盘中的存储路径为“data\chap3\example3_1”。请分性别分析学生的英语成绩分布情况。

操作步骤：

步骤一：选择菜单 Solutions|ASSIST 进入 ASSIST 分析主界面，选择菜单 Data Analysis|Elementary|Summary Statistics 进入如图 3-4 所示对话框。

步骤二：单击 Table 按钮，在弹出的对话框中单击 libraries 下的 chap3 并单击在 Table（数据集）下随之出现的 example3_1，单击 OK 按钮完成选定分析数据集 chap3.example3_1。单击 Subset data 按钮，在弹出的对话框（如图 3-5 所示）中单击按钮 BY columns，在弹出的对话框中选定分组变量 sex。

VIEWTABLE: Chap3.Example3_1

	id	sex	score
1	1	1	77
2	2	0	79
3	3	0	77
4	4	0	71
5	5	0	78
6	6	0	54
7	7	1	78
8	8	0	74
9	9	1	80
10	10	0	86

图 3-3　chap3.example3_1 部分数据

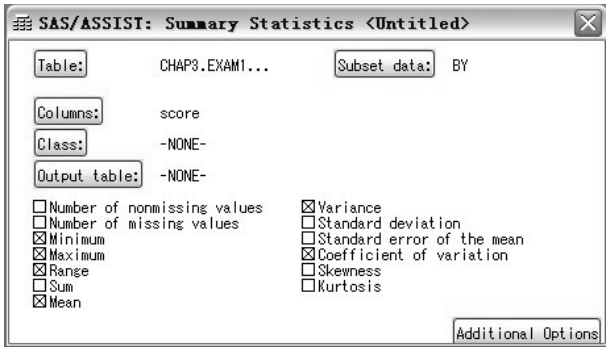


图 3-4　Summary Statistics 对话框

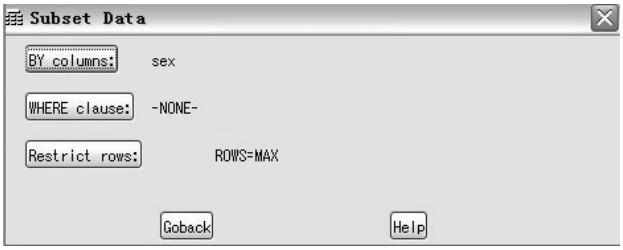


图 3-5　设定分析限制条件

注意：在设定分类变量时需要对数据集事先进行排序，因此单击 Sort Data 按钮，弹出对话框如图 3-6 所示，单击 Sort by columns 按钮，在弹出的对话框中选定分类变量 sex，选择菜单 Run|Submit 完成排序。单击提示排序成功的对话框上的 OK 按钮，返回如图 3-5 所示对话框；单击 OK|Goback 返回如图 3-4 所示对话框。

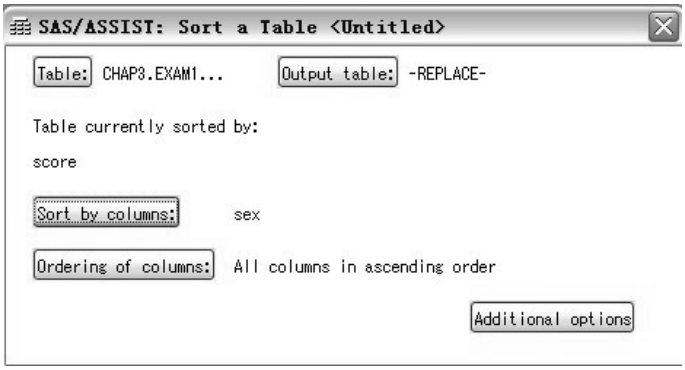


图 3-6　对数据集排序

步骤三：单击 Column 按钮，在弹出的对话框中选定分析变量 score。在 Output Table 下列出的一些指标中以单击选项前空格的方式设定计算输出 Number of nonmissing values（非缺失样本数）、Minimum（最小值）、Maximum（最大值）、Mean（均值）、Variance（方差）、Coefficient of Variation（变异系数）。选择 Run|Submit 菜单提交设置，则在 Output 输出记录窗口显示的结果如图 3-7 所示：分性别计算了学生的最低成绩、最高成绩、平均成绩、成绩的方差和变异系数。

注意：在 Log 输出记录窗口将有完成菜单操作的相应 SAS 程序，如图 3-8 的 331～340 行所

示，本操作调用了 means 过程完成分析。建议对 SAS 系统不够熟练的用户充分利用此信息学会编写自己的 SAS 程序。

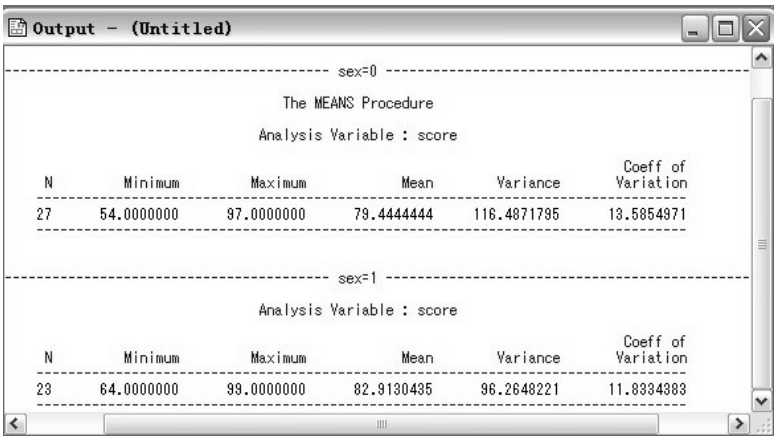


图 3-7　结果输出

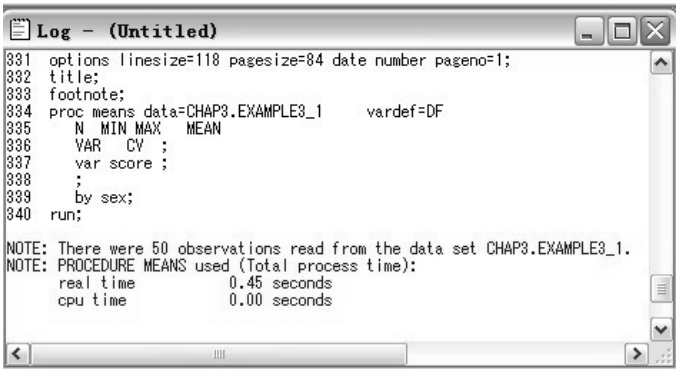


图 3-8　Log 记录窗口记录操作程序

3.2　SAS/INSIGHT 交互分析介绍

3.2.1　SAS/INSIGHT 概述

选择菜单 Solutions|Analysis|Interactive Data Analysis，打开如图 3-9 所示对话框。此时需选定进入分析的数据集，如单击 chap3，再单击 example3_1，单击 Open 按钮打开此数据集并进入分析界面。此时可用的菜单栏为 File、Edit 和 Analyze，以下分别介绍其功能。

File：新建、打开、保存、打印文件等功能。

Edit：设定界面（Window 子菜单）、转换变量（应用 Variables 子菜单实现倒数变换、对数变量、开平方、指数变换等）、有关观测的操作（查找观测、在图形和分析中显示/隐藏特定观测等）。

Analyze：作为 INSIGHT 最重要的分析菜单，通过选择其子菜单可以对应实现的绘图和分析功能如图 3-10 所示。

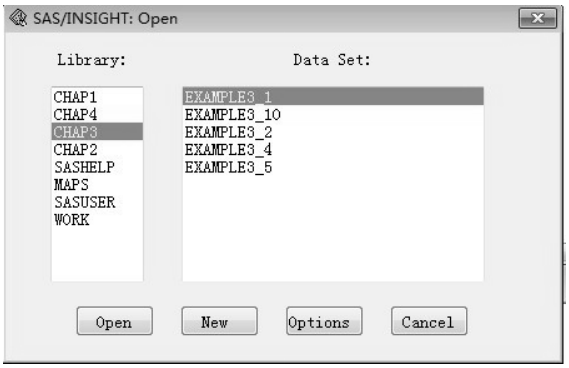


图 3-9　打开数据集

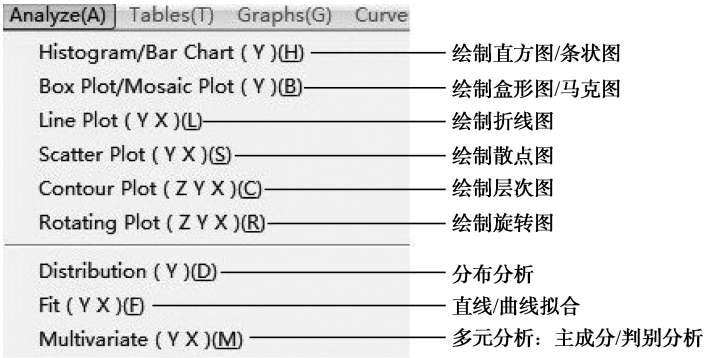


图 3-10　Analyze 菜单的子菜单

3.2.2　SAS 实例——绘制身高和体重的散点图

例 3-2　已知在某班随机抽查了 10 名女生，并测量记录其身高和体重数据保存在数据集 chap3.example3_2 中（如图 3-11 所示），该数据集在光盘中的存储路径为“data\chap3\example3_2”。请用 SAS/INSIGHT 模块以菜单操作的方式绘制关于女生身高和体重的散点图。

VIEWTABLE: Chap3.Example3_2

	学号	身高	体重
1	01	169.5	71
2	09	166	58
3	12	157	56.5
4	14	180	64.6
5	23	158	53
6	45	165	51
7	16	164	56
8	18	158	49
9	32	168	55
10	28	167	71

图 3-11　数据集 chap3.example3_2

操作步骤：

步骤一：选择菜单 Solutions|Analysis|Interactive Data Analysis，在弹出的对话框中单击选择数据集 chap3.example3_2，单击 Open 按钮打开数据集，进入分析界面。

步骤二：选择菜单 Analyze|Scatter Plot（Y,X），在弹出对话框中单击变量 height，再单击 Y 按钮，设置 height 为纵轴变量；单击变量 weight，再单击 X 按钮，设置 weight 为横轴变量。单击 OK 按钮，则输出散点图如图 3-12 所示。

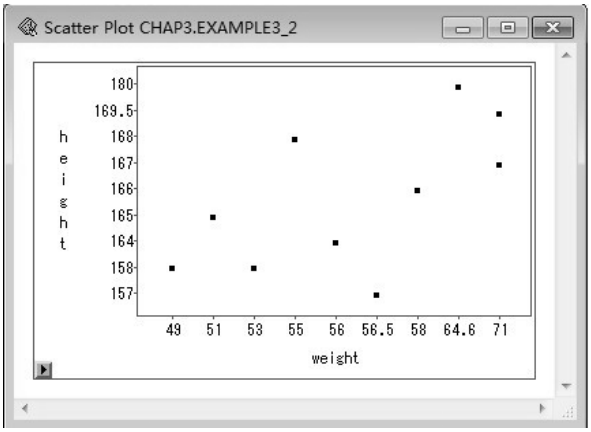


图 3-12　weight*height 散点图

3.3　Analyst（分析家）模块操作

3.3.1　Analyst 模块概述

Analyst 模块能实现较为完善的数据整理和统计分析功能，能适应不同层次用户的需求。而且每次分析都会形成一个项目文件，包含菜单分析的所有代码，便于用户完善和修改程序、学习 SAS 编程。

选择 Solutions|Analysis|Analyst 菜单进入如图 3-13 所示 Analyst 模块主界面。Analyst 的分析功能主要通过选择菜单栏上的选项完成，Analyst:（new project）下的窗口被分成了左侧的目录树和右侧的数据显示窗口。

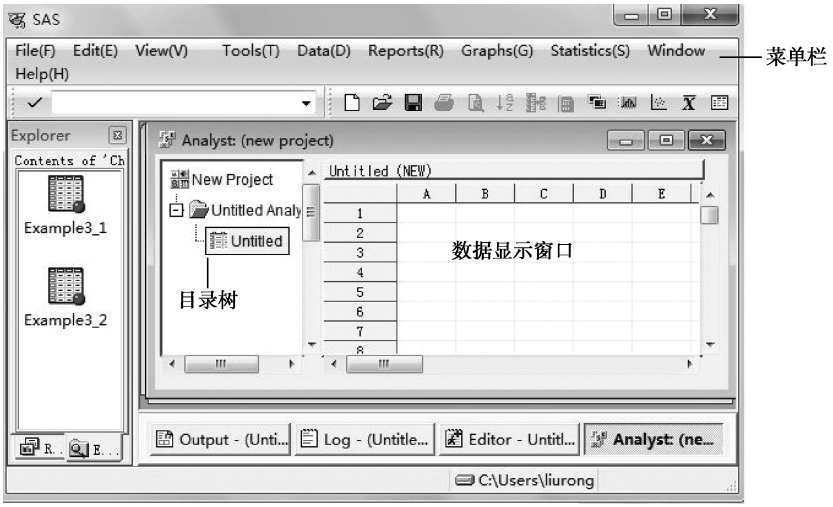


图 3-13　Analyst 模块主界面

菜单栏上的选项（从左至右）主要实现的功能如下。

File：新建、打开、保存数据集，退出 Analyst 模块。

Edit：编辑数据，切换数据集的浏览和编辑模式。

View：可视化变量，主要包括移动、隐藏、显示变量等。

Tools：系统全局变量设置，包括输出的标题、视图的设置等。

Data：数据管理，具体内容将在 3.3.2 节介绍。

Reports：生成报表。

Graphs：绘制图形，主要有条形图、饼图、盒形图、散点图、正态概率图（P-P 图）、层次图和曲面图。

Statistics：统计分析，具体内容将在 3.3.3 节介绍。

Window：界面管理。

Help：进入帮助系统。

3.3.2 应用 Analyst 整理数据

本节将以例题的方式逐一介绍应用 Data 菜单实现各项数据管理。

功能一：筛选观测

例 3-3 筛选数据集 chap3.example3_2 中身高大于 165cm 的女生，将数据集另存为 chap3.example3_3。

操作步骤：

步骤一：选择菜单 File|Open by SAS name，在弹出的对话框中单击打开数据集 chap3.example3_2。

步骤二：选择菜单 Data|Filter|Subset Data，弹出如图 3-14 所示对话框，单击变量 height，则 Operators（操作符）按钮下出现操作运算符下拉菜单：EQ（等于）、NE（不等于）、GT（大于）、LT（小于）、GE（大于等于）、LE（小于等于）、||.Concatenate（连接字符串）、IN（在某指定范围中间）、Not IN（不属于某范围）、OTHER Operators（其他不常见的操作符）。本例单击选择 GT，则此运算符出现在 Where 选项框内。

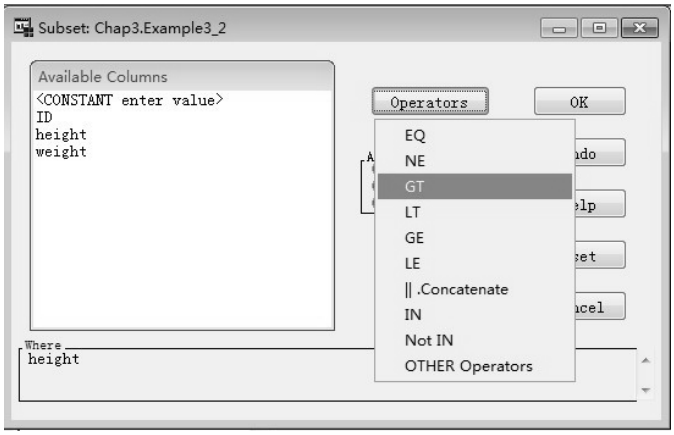


图 3-14 Analyst/筛选观测

步骤三：单击 CONSTANT enter value，在弹出的对话框内输入 165，单击 OK 按钮保存设置并返回如图 3-14 所示的对话框。单击 OK 按钮，则数据集 chap3.example3_1 仅保存了 height 大于 165 的观测。

步骤四：选择菜单 File|Save As by SAS name，在弹出的对话框中指定将当前数据集存储在逻辑库 chap3 中，并在光标闪烁的空白栏输入数据集名 example3_4，单击 Save 按钮完成设置。

注意：接下来所有的操作，请注意选择菜单 Edit|Mode|Edit Mode 设置编辑模式，此处设置非常重要，否则在系统默认的 Browse（浏览）模式下几乎所有的数据管理功能实现的菜单都将显示不可用的灰色。

功能二：数据排序

例 3-4　请对数据集 chap13.example3_2 按照变量 weight 的值升序排列。

操作步骤：

步骤一：选择菜单 File|Open by SAS name，在弹出的对话框中单击打开数据集 chap3.example3_2。选择菜单 Edit|Mode|Edit Mode 设置编辑模式。

步骤二：选择菜单 Data|Sort，弹出如图 3-15 所示对话框，单击选择变量 weight，然后单击 OK 按钮即可。

注意：此处只可设置升序排列，若用户需要设置降序排列，可用第 2 章介绍过的 PROC SORT 过程实现。

功能三：数据转换

例 3-5　根据数据集 chap3.example3_2 中的信息，新增变量 BMI（BMI=weight/(height)**2，其中 height 的单位为 m，weight 的单位为 kg）和变量 level，若 BMI<18.5 则 level 取值为 1；若 18.5≤BMI<24，则 level 取值为 2；若 24≤BMI<28，则 level 取值为 3；若 BMI≥28，则 level 的取值为 4，将数据集另存为 chap3.example3_5。

操作步骤：

步骤一：选择菜单 File|Open by SAS name，在弹出的对话框中单击打开数据集 chap3.example3_2。选择菜单 Edit|Mode|Edit Mode 设置编辑模式。

步骤二：首先计算 BMI 的值。选择菜单 Data|Transform|Compute 打开如图 3-15 所示对话框。删除 comp1 并在此内输入计算变量 BMI。在计算框内输入 BMI 的计算公式 weight/(height/100)**2。注意，需要将数据集 chap3.example3_2 中的变量 height 从 cm 转换成 m。单击 OK 按钮完成计算。

注意：在计算时可以应用 Category（类别）下的各类计算公式，只需要选定相应的计算公式，再单击其上按钮▲即可将其选入计算框。

步骤三：根据 BMI 的值新建等级变量 level。选择菜单 Data|Transform|Record Ranges 打开如图 3-16 所示对话框。在 Column to recode 选项后面输入变量 BMI，然后将 New column name 选项后的 BMI_recode 更改为 level。新建变量属性（New column type）采用默认的 Numeric（数值型变量）。在 Number of groups to be formed（新建变量组数）中填入 4。单击 OK 按钮进入如图 3-17 所示对话框。首先选择 Operators 选项下的“<=and<”单选按钮，然后设置 BMI 的范围及对应变量 level 的取值。单击 OK 按钮完成设置。

步骤四：选择菜单 File|Save As by SAS name，在弹出的对话框中指定存储逻辑库 chap3，输入数据集名 example3_5，单击 Save 按钮完成设置。

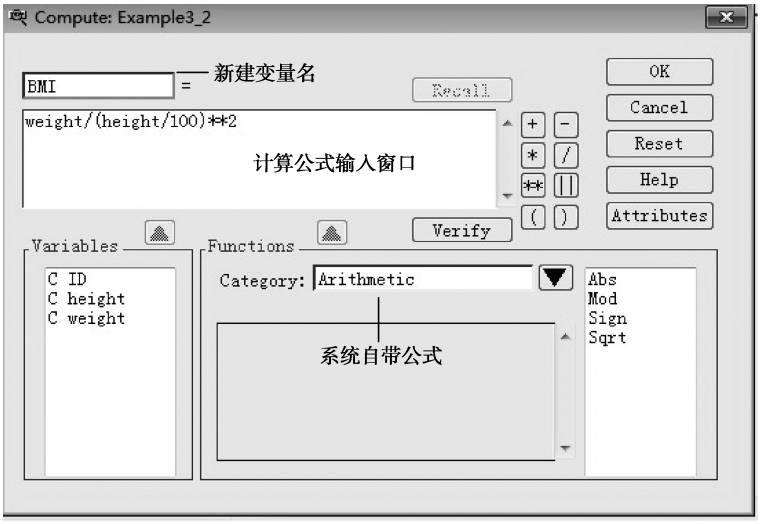


图 3-15　计算变量

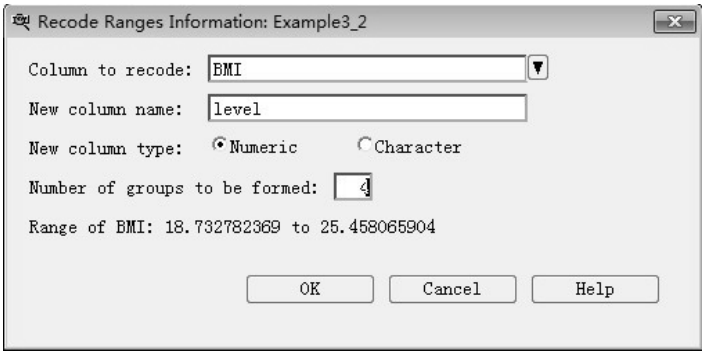


图 3-16　设置新建变量名

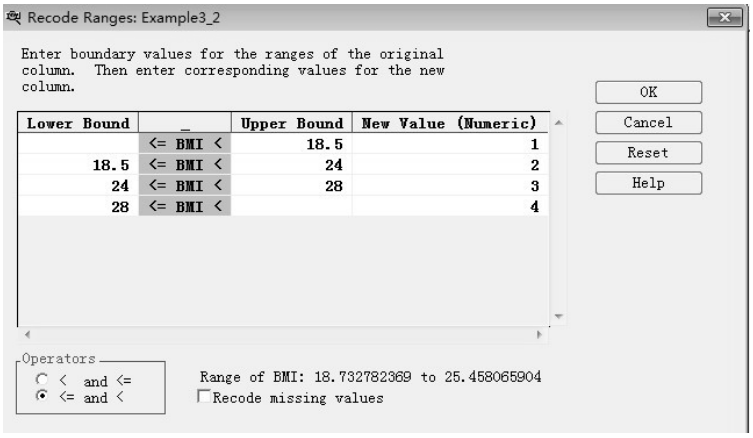


图 3-17　设置新建变量取值

注意：在 Data|Transpose 菜单下，还可以实现将数据转换成秩（Rank）、数据标准化（Standardize）、Recode Value（根据离散变量的取值新建编码新变量）、Covert Type（转换数据类型）。希望读者在学习了例 3-5 后能够举一反三，遇到实际问题时灵活使用数据转换功能。

功能四：产生随机数

例 3-6　在数据集 chap3.example3_5 中新建变量 score，其取值为一组符合均值为 80 的泊松分布的随机数，并另存为数据集 chap3.example3_6。

操作步骤：

步骤一：选择菜单 File|Open by SAS name，在弹出的对话框中单击打开数据集 chap3.example3_5。选择菜单 Edit|Mode|Edit Mode 设置编辑模式。

步骤二：选择菜单 Data|Random Variables|Possion 打开如图 3-18 所示对话框。删除 New column name（新变量名）后系统默认的新建变量 Possion1，并输入 score，在 Parameter（参数）选项框的 Mean（均值）后填入 80，单击 OK 按钮。

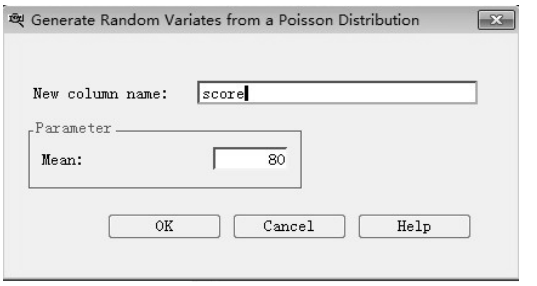


图 3-18　设置分布参数

步骤三：选择菜单 File|Save As by SAS name，在弹出的对话框中指定存储逻辑库 chap3，输入数据集名 example3_6，单击 Save 按钮完成设置。

注：在 Random Variables 后选择的分布具体有：

- Normal——正态分布
- Uniform——标准正态分布
- Biomial——二项分布
- Chi-Sqaure——卡方分布
- Possion——泊松分布
- Beta——贝塔分布
- Exponential——指数分布
- Gamma——伽马分布
- Genometric——几何分布
- Extreme Value——极值

功能五：分组描述统计

例 3-7　基于数据集 chap3.example3_6 按照变量 level 分组计算模拟的变量 score 的均值和标准差。

操作步骤：

步骤一：选择菜单 File|Open by SAS name，在弹出的对话框中单击打开数据集 chap3.example3_6。选择菜单 Edit|Mode|Edit Mode 设置编辑模式。

步骤二：选择菜单 Data|Summarize by Groups 打开如图 3-19 所示对话框。将变量 level 选入 Group 选项框，变量 score 选入 Summarize 选项框内。选择 Statistics（统计量）选项框内 Mean（均值）和 Standard deviation（标准差）复选框，设置计算这两个统计量。单击 OK 按钮即可得

到结果数据集 Example3_2 Summarized by Group Table（如图 3-20 所示）。由此可知全部样本的 score 均值为 80，标准差为 11.3（保留一位小数）；level 为 2 和 3 的样本 score 均值和标准差在数据集第二、第三行读取。

注意：score 的值由例 3-6 模拟得到，它的每一次模拟取值都不一样，因此本例据此得到的结果也有细微差别。

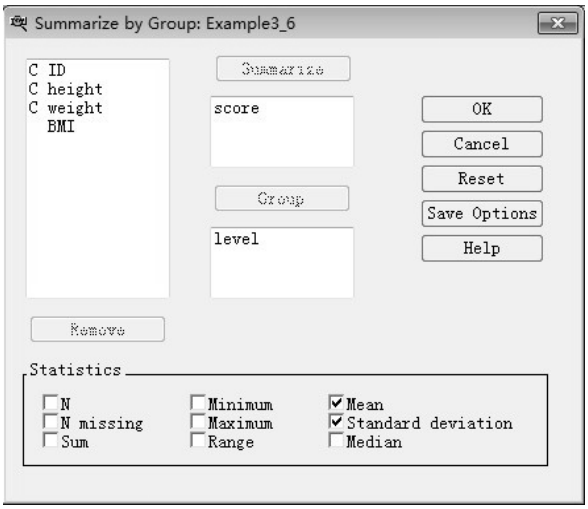


图 3-19　分组描述统计

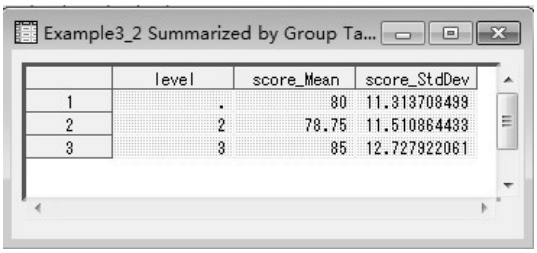


图 3-20　结果数据集

功能六：数据转置

例 3-8　请将数据集 chap3.example3_5 按照变量 level 分类进行转置。

操作步骤：

步骤一：选择菜单 File|Open by SAS name，在弹出的对话框中单击打开数据集 chap3.example3_5。选择菜单 Edit|Mode|Edit Mode 设置编辑模式。

步骤二：选择菜单 Data|Transpose 进入如图 3-21 所示对话框，单击选择变量 height、weight 和 BMI，再单击按钮 Transpose，设置对以上变量进行转置。类似的，将分组变量 level 选入 Group By 选项框。单击 OK 按钮完成设置。

功能七：随机抽样

例 3-9　在数据集 chap2.example3_2 中随机抽取 4 条观测。

操作步骤：

步骤一：选择菜单 File|Open by SAS name，在弹出的对话框中单击打开数据集 chap3.example3_2。选择菜单 Edit|Mode|Edit Mode 设置编辑模式。

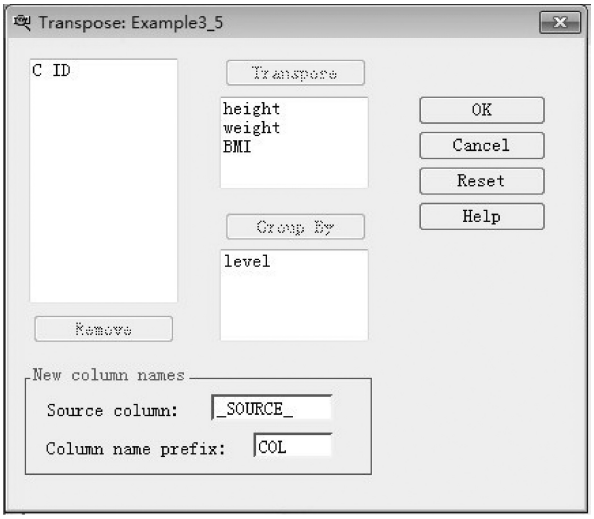


图 3-21　转置数据集

步骤二：选择菜单 Data|Random Sample 打开如图 3-22 所示对话框，在此可以设定随机抽取样本的个数（Rows），则 Ratio（比率）后的参数会相应的调整。同样可以设置抽样比率（Ratio）。本例在 Rows 后填入 4，单击 OK 按钮完成设置。

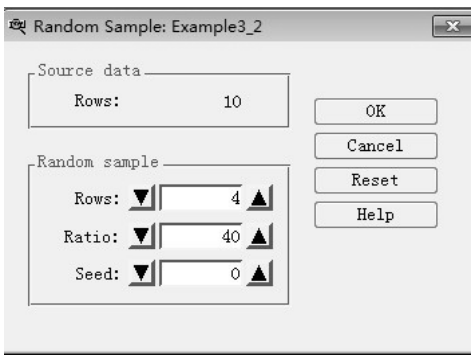


图 3-22　随机抽样

除了以上介绍的常用功能，还可应用 Analyst 的 Data 子菜单下的 Combine Table（合并数据集）实现数据集的纵向连接（Merge by Columns）和横向合并（Concatenate by rows）；Stack column 实现变量的堆栈；Split column 根据变量拆分数据集。

3.3.3　应用 Analyst 进行统计分析

本节将概述 Analyst 模块能够实现的统计分析，不同分析的具体操作将在以后的章节中详细介绍。

选择菜单 Statistics，出现的下拉菜单中的选项依次如下。

➢ Descriptive：描述性统计分析

- Summary statistics——描述性统计指标
- Distributions——数据分布

- Correlation——相关分析
- Frequency Counts——频数分析

➢ Table Analysis：列联表分析

➢ Hypothesis test：假设检验

- One-Sample Z-test for Mean——单样本 *Z* 检验
- One-Sample t-test for Mean——单样本 *t* 检验
- One-Sample Test for proportion——单样本比率检验
- One-Sample Test for the Variance——单样本方差检验
- Two-Sample t-test for Means——两样本 *t* 检验
- Two-Sample Paired t-test for Means——配对样本 *t* 检验
- Two-Sample Test for proportion——两样本比率检验
- Two-Sample Test for the Variance——两样本方差检验

➢ ANOVA：方差分析

- One-Way ANOVA——单因素方差分析
- Non-parametric One-Way ANOVA——非参数单因素方差分析
- Factorial ANOVA——析因设计方差分析
- Linear Model——一般线性模型
- Repeated Measures——重测数据的方差分析
- Mixed Model——混合模型方差分析

➢ Regression：回归分析

- Simple——一元回归（包括直线、抛物线和三次曲线）
- Linear——多元线性回归
- Logistic——逻辑斯蒂回归

➢ Multivariate：多元分析

- Principal Components——主成分分析
- Canonical Correlation——典型相关分析

➢ Survival：生存分析

- Life Tables——生命表分析
- Proportional Hazards——相对危险率分析

➢ Sample Size：样本量估计

其主要为在不同的统计模型背景下，根据样本量范围计算相应的统计功效，或者反之，根据设定的统计功效计算达到此统计功效所需要的样本数。

3.3.4 SAS 实例——探索年龄和血压的相关关系

例 3-10 已知数据集 chap3.example3_10 （在光盘中的存储位置为“data\chap3\example3_10”）中包含了抽查的 15 个成人男性的收缩血压和年龄数据。请应用 SAS/ANALYST 相关模块以菜单操作的方式分别计算年龄和收缩血压的相关系数。

操作步骤：

步骤一：选择菜单 Solutions|Analysis|Analyst 进入 Analyst 主界面。选择菜单 File|Open by SAS Name|chap3|example3_10，单击 OK 按钮打开数据集 chap3.example3_10。

步骤二：选择菜单 Statistics|Descriptions|Correlations 打开如图 3-23 所示对话框。选择变量 age（年龄）和变量 SBP（收缩血压），再单击 Correlate（相关）按钮，设置计算这两个变量的相关系数。单击 OK 按钮完成分析。根据 Output 结果输出窗口的显示（如图 3-24 所示）可知年龄和收缩血压的相关系数为 0.80888。

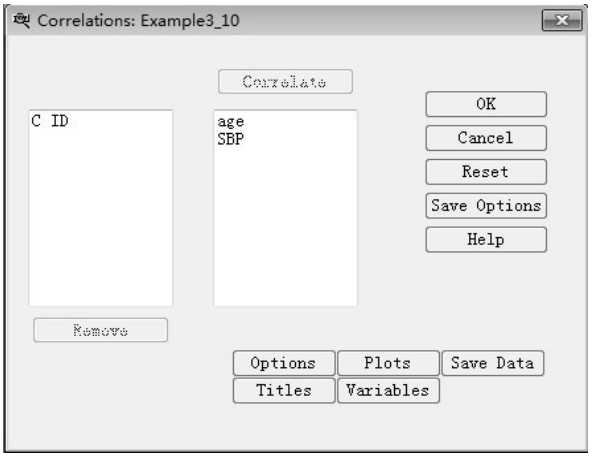


图 3-23　相关分析

Pearson Correlation Coefficients, N = 20
Prob > |r| under H0: Rho=0

	age	SBP
age	1.00000	0.80888 <.0001
SBP	0.80888 <.0001	1.00000

图 3-24　相关系数计算结果

注意：在完成分析后 Analyst 模块左界面的树状目录中 Correlations 下的 Correlation of Example3_8 为输出的结果，而 Code 为实现分析的 SAS 程序，读者可通过阅读此类 Code 来学习和强化 SAS 编程能力。

练习题

习题 3-1　已知统计某院教师的年龄和性别信息如表 3-1 所示，相应的 SAS 数据集在光盘中的存储路径为“data\chap3\exercise3_1”。

（1）调用 ASSIST 模块分性别计算教师年龄的均值和标准差；

（2）调用 Analyst 模块分析教师年龄的分布，并绘制直方图；

（3）调用 INSIGHT 模块分析教师年龄的分布，并绘制直方图。

表 3-1　某院教职工年龄统计表

编号（ID）	年龄（age）	性别（sex）	编号（ID）	年龄（age）	性别（sex）	编号（ID）	年龄（age）	性别（sex）
1	28	男	20	29	男	40	52	女
2	44	男	21	54	女	41	52	女
3	39	男	22	57	女	42	72	男
4	43	女	23	38	女	43	72	女
5	37	女	24	59	女	44	36	男
6	50	男	25	34	女	45	52	男
7	32	男	26	96	男	46	34	男
8	38	男	27	40	男	47	38	男
9	32	女	28	44	男	48	64	女
10	33	男	29	29	男	49	41	男
11	48	女	30	41	女	50	46	女
12	33	女	31	76	女	51	19	女
13	28	女	32	67	女	52	59	女
14	4	女	33	54	男	53	41	男
15	30	男	34	60	男	54	56	女
16	53	男	35	47	女	55	43	女
17	53	男	36	46	女	56	32	女
18	41	女	37	19	女			
19	58	男	38	51	男			

习题 3-2　已知条件同习题 2-2，SAS 数据集为“data\chap2\exercise2_2”。请分别应用 INSIGHT 模块、Analyst 模块和 INSIGHT 模块以菜单操作的方式完成以下分析：

（1）请计算应聘者的加权得分，其中专业技能的权重为 0.4，言行举止的权重为 0.3，思维活跃度的权重为 0.3；

（2）请按照如下规则根据应聘者的加权得分评定其等级：加权得分大于 90 分——“A”；加权得分在 80～90 分——“B”级；加权得分在 70～80 分——“C”级；加权得分低于 70 分——“D”级。

第 4 章　定量数据描述性统计分析

以上章节概述了 SAS 系统基础及 SAS 编程语言，应用这些知识可建立 SAS 数据集并进行简单的数据预处理。此后的章节将重点介绍统计分析。本章主要介绍应用 SAS/BASE 中的 MEANS、UNIVARIATE 过程对连续型数据进行描述性统计分析（有关离散型数据的分析将在“列联分析”章节中详细介绍），以及利用 SAS/GRAPH 软件模块的 GPLOT 图形过程和 GCHART 图表过程绘制直方图、散点图、条形图等常用统计图形。

4.1　描述性统计分析指标

描述性统计分析指标主要用于描述数据的分布情况，一般被分为描述数据集中位置和离散程度两类指标。本节将详细介绍常见描述性统计分析指标的计算公式和意义，并引入实例计算相应的统计分析指标。

4.1.1　基本指标介绍

描述性统计分析指标基于以下基本统计学概念：

总体：研究对象的全体。

样本：从总体中抽取的个体。

随机抽样：从总体中抽取样本，且每个样本被抽到的机会均等。

例如，为了解某学院大一新生的平均年龄，用计算机产生随机数的方式抽取 50 名学生的学号，计算出他们的年龄（精确到天）及其平均值。以上随机抽样过程，总体是某学院大一新生的年龄，样本是抽取的 50 名学生的年龄，样本量即为 50，采取了随机抽样方式，即每个学生被抽到的机会均等。

描述性统计指标可被分为描述数据集中位置和离散程度两类，描述数据集中位置的有均值、中位数、众数等；描述数据离散程度的主要有方差、标准差、变异系数等。以下列出了主要描述性统计指标的名称、数学表达式及意义。

- MEAN（均值）：计算公式为 $\text{mean}=\left(\sum_{i=1}^{n}x_i\right)/n$ （n 为样本量 N）。
- MODE（众数）：样本中出现次数最多的数据。
- MEDIAN（中位数）：指将数据按大小顺序排列起来形成一个数列，居于数列中间位置的数据。若总数为奇数，取中间值；若总数为偶数，取中间两个值的平均值。
- P_X（分位数）：它将全部观察值分成两个部分，其中有 x% 个观察值小于 P_X，（100−x）% 个观察值大于 P_X。

- MAX（最大值）：样本中的最大观察值。
- MIN（最小值）：样本中的最小观察值。
- SUM（和）：样本观察值的总和。
- RANGE（极差）：最大与最小观测值之差。
- STD DEV（标准差）：计算公式为STD　DEV $=\sqrt{\dfrac{\sum(x-\bar{x})^2}{n-1}}$（$n$ 为样本量 N）。
- VAR（方差）：为标准差的平方，用来衡量相对于均值的分散性和变异性。数据的集中程度高则方差小，反之则大。
- STDERR（标准误）：计算公式为 $\text{STDERR}=\text{STD}/\sqrt{N}$ 。
- CV（变异系数）：计算公式为 $\text{CV}=\dfrac{s}{\bar{x}}\times 100\%$ 。
- USS（加权平方和）：计算公式为 $\text{USS}=\sum_{i=1}^{n} w_i x_i^2$ ，其中 w_i 代表权重。
- CSS（加权离差平方和）：计算公式为 $\text{CSS}=\sum_{i=1}^{n} w_i(x_i-\bar{x})^2$ 。
- SKEWNESS（偏度系数）：计算公式为 $\text{SKEW}=\dfrac{n}{(n-1)(n-2)}\sum\left(\dfrac{x_i-\bar{x}}{s}\right)^3$，主要衡量数据的对称性，若其值大于 0 则表示位于均值右边数据较分散，若小于 0 则表示均值左边数据较分散。
- KURTOSIS（峰度系数）：计算公式为 $\text{KURT}=\dfrac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\dfrac{x_i-\bar{x}}{s}\right)^4-\dfrac{3(n-1)^2}{(n-2)(n-3)}$，峰值反映分布的尖锐度或平坦度，正峰值表示相对尖锐的分布，负峰值表示相对平坦的分布。
- PROB>|T|：在总体均值是 0 的假设条件下，学生 t 统计量大于临界 T 的绝对值的概率。

以上指标中较易混淆的是标准差（STD DEV）和标准误（STDERR）。两者的区别在于：标准差衡量样本观察值的离散程度，而标准误衡量根据样本计算的统计量的标准差，如在参数估计时得到的对应的标准误衡量的是参数估计值和实际值的差异。标准差的值越大表示观察值的分布越分散；反之，标准差越小表示观察值的分布越集中。标准误是统计量的标准差。由于统计量是样本观察值的函数，一旦样本改变则统计量的取值也随之改变。在参数估计中，用样本的统计量去估计参数时，统计量的标准误越小，表示抽样误差小、统计量较稳定、与参数较接近。

4.1.2　SAS 过程——MEANS 过程

MEANS 过程主要用于连续型数值型变量中产生针对单个变量的描述性统计量，在没有指定输出统计量时，系统默认输出 N（样本量）、MEAN（均值）、STD DEV（标准差）、MIN（最小值）、MAX（最大值）5 类统计量。MEANS 过程语句使用格式如下：

```
PROC MEANS DATA=SAS 数据集 <选项列表> ;
VAR       变量列表;
CLASS     变量列表;
BY        变量列表;
FREQ      变量;
WEIGHT    变量;
ID        变量列表;
OUTPUT    <OUT=输出数据集名>   <统计量关键字=变量名列表>;
RUN;
```

PROC MEANS 语句后的<选项列表>主要内容如下：

VARDEF=DF/WEIGHT/WGT/N/WDF——指定方差的计算公式中的除数 D。若 VARDF= DF 则 $D=N-1$（系统默认）；若 VARDF=WEIGHT/WGT 则 $D=\sum W_i$；若 VARDF=N 则 $D=N$；若 VARDEF=WDF 则 $D=\sum W_i-1$。

NOPRINT——不输出任何描述性统计量。

MAXDEC=数字——输出结果中小数点位数（0～8）（系统默认值为 8）。

DESCENDING——指定输出的数据集按照_TYPE_值降序排列（系统默认升序）。

ALPHA=数字——设置计算置信区间的置信水平 α，α 值为 0～1。

统计量——指定输出分析指标，可使用的关键字如表 4-1 所示。

表 4-1　MEANS 过程可选统计量关键字

统 计 量	含　义	统 计 量	含　义
N	不包含缺失值的观测数目	MODE	众数，出现频数最高的数
NMISS	包含缺失值的观测数目	SUMWGT	加权数和
MEAN	平均数	MAX	最大值
STDERR	均值的标准误	MIN	最小值
SUM	加权和	RANGE	极差，最大值减去最小值
STD	标准差	MEDIAN	中位数
VAR	方差	T	总体均值等于 0 的 t 统计量
CV	变异系数	PRT	t 分布的双尾 P 值
USS	加权平方和	CLM	置信上限和下限
CSS	均值偏差的加权平方和	LCLM	置信下限
SKEWNESS	对称性的度量——偏度	UCLM	置信上限
KURTOSIS	尾部陡平度的度量——峰度		

MEANS 过程所使用的语句意义如下：

VAR 语句——指定进行描述性统计分析的变量，并指定变量输出顺序，如语句“var height weight;”则先输出变量 height 的结果，再输出变量 weight 的结果。

BY 语句——指定分组变量，按 BY 语句定义的变量分组计算其相应的统计量，注意使用前应先按 BY 变量对数据集排序，否则系统将报错。

CLASS 语句——定义观测组，分组计算观测的统计量。

FREQ 语句——指定一个数值型的频数变量，它的值表示输入数据集中相应观测出现的频数。

WEIGHT 语句——指定加权变量，它的值表示相应观测的权数。

ID 语句——为识别输出数据集里的观测，在输出数据集中增加的一个或几个附加变量。

在 MEANS 过程，使用 BY 或 CLASS 语句得到的结果意义是一样的，但使用 BY 语句将按照 BY 变量的不同取值分别输出多张结果报表，而使用 CLASS 语句输出结果于一张报表，不同行代表 CLASS 语句的不同取值。MEANS 过程对 OUTPUT 语句的次数没有限制，可使用多个 OUTPUT 语句来创建内容不同的多个数据集，OUTPUT 语句后的选项意义如下：

<OUT=输出数据集名>——输出数据集名。

统计量关键字=变量名列表——指定输出的数据集中包含的统计量与它们在新数据集中的变量名。可定义输出的统计量关键字如表 4-1 所示。例如，使用语句“output out=chap4.test mean=sample_mean”定义将计算结果 mean 保存到数据集 chap4.test 中，且将 mean 重命名为 sample_mean。

4.1.3 SAS 过程——UNIVARIATE 过程

SAS 系统的 UNIVARIATE（单变量）过程主要用于对指定随机变量进行详细的描述性统计，不仅包含 MEANS 过程的功能，还可计算一些其他的统计量并生成统计图（茎叶图、盒形图和正态概率图）。以下简介此过程输出的统计图。

茎叶图（STEM-AND-LEAF DISPLAY）：用于形象地初步描述数据分布，每一数据被分成茎、叶和可以忽略部分进行描述，类似直方图。

盒形图（BOXPLOT）：由一个矩形和两条平行线组成，上线为 75%分位数，下线为 25%分位数，两条线之间的（+）号标识出平均数。矩形盒较短表明数据比较集中；两端的触须线对称或长短不一反映数据的分布特性。

正态概率图（NORMAL Q-Q PLOT）：主要用于辅助判断数据是否服从正态分布。它以实际观测值为纵轴、以标准百分位的百分位数为横轴，在图中用（*）号代表实际观测值，用（+）号标识一条根据数据平均数与标准差画出的参考线，若观测值服从正态分布，则星号（*）落在加号（+）上，即两者重叠多。

UNIVARIATE 过程的一般使用格式如下：

```
PROC UNIVARIATE DATA=SAS 数据集 <选项列表> ;
VAR        变量列表;
BY         变量列表;
FREQ       变量;
WEIGHT     变量;
ID         变量列表;
OUTPUT   <OUT=输出数据集名> <统计量关键字=变量名列表>
<PCTLPTS=百分位数 PCTLPRE=变量前缀名 PCTLNAME=变量后缀名>;
RUN;
```

PROC UNIVARIATE 语句后的<选项列表>主要内容如下：

VARDEF=DF/WEIGHT/WGT/N/WDF——指定方差计算中的除数 D，取值的意义同 MEANS 过程。

FREQ——要求生成包括变量值、频数、百分数和累计频数的频率表。

NORMAL——要求计算关于输入数据服从正态分布的假设的检验统计量。

PLOT——要求生成茎叶图、盒形图和正态概率图。

ROUND=舍入单位列表——指定 VAR 语句中变量的四舍五入的单位。

指定输出统计量的关键字，如表 4-2 所示。

表 4-2　UNIVERIATE 过程中的统计量关键字

统 计 量	含　义	统 计 量	含　义
N	观测数目	MODE	众数，出现次数最多的数
NMISS	包含缺失值的观测数目	T	总体均值等于 0 的 *t* 统计量
NOBS	观测个数	PRT	*t* 分布的双尾 *P* 值
MEAN	算术平均值	Q3	上四分位数（75%）
STDERR	均值的标准误	Q1	下四分位数（75%）
SUM	加权和	QRANGE	上下四分位数差（Q3−Q1）
STD	标准差	P1	1%分位数
VAR	方差	P5	5%分位数
CV	变异系数	P10	10%分位数
USS	加权平方和	P90	90%分位数
CSS	均值偏差的加权平方和	P95	95%分位数
SKEWNESS	对称性的度量——偏度	P99	99%分位数
KURTOSIS	尾部陡平度的度量——峰度	MSIGN	符号统计量
SUMWGT	加权数和	PROBM	大于符号秩统计量的绝对值概率
MAX	最大值	SIGNRANK	符号秩统计量
MIN	最小值	PROBS	大于中心符号秩统计量的绝对值 *P*
RANGE	极差，最大值减去最小值	NORMAL	检验正态分布的统计量
MEDIAN	中位数	PROBN	检验正态分布假设的概率值

OUTPUT 语句中主要选项如下：

<PCTLPTS=百分位数 PCTLPRE=变量前缀名 PCTLNAME=变量后缀名>——提供自定义计算的百分位数和指定其在输出数据集中合成的变量名。

统计量关键字=变量名列表——指定在输出数据集中要包含的统计量并将这些统计量在新数据集重命名。

UNIVARIATE 过程可使用 VAR、BY、FREQ、WEIGHT、ID 语句，这些语句的用法与意义和 MEANS 过程完全一样，不再赘述。

4.1.4　SAS 实例——描述小麦单穗粒数分布

例 4-1　在某农业试验基地进行试验，从某块农田中随机抽取 50 株小麦，并且测出其单穗粒数（count）如表 4-3 所示（相应的 SAS 数据集在光盘中的存储路径为“data/chap4/wheat”），

请据此计算其描述性统计指标，以得到小麦单穗粒数的分布状况。

表 4-3　小麦单穗粒数

29	26	34	25	36	31	32	22	43	29
29	30	33	28	29	22	27	33	32	29
25	25	36	31	27	32	26	29	21	25
27	27	30	26	27	18	29	28	30	27
32	24	28	28	34	25	27	26	25	32

解析：此实例不仅可采用 MEANS 或 UNIVARIATE 过程以编程的方式解答，还可以运用 SAS/ANALYST、SAS/INSIGHT 和 SAS/ASSIST 以菜单操作的方式分析。本例将以编程和 SAS/ANALYSIT 菜单操作完成分析，运用其他两种模块以菜单操作方式进行描述统计分析请读者参考第 3 章菜单操作的相关小节自行分析。

编程法：

编写程序如下所示，其在光盘中的存储路径为“proc/chap4/wheat”。请读者注意从本章开始将全部应用 ODS 输出系统将结果以 RTF 文本格式的形式输出，并逐条解释结果。

```
ods listing close;                          /*关闭列表输出*/
ods rtf file= 'chxy.rtf';                   /*指定结果以 RTF 文本输出*/
proc means data=chap4.wheat MAXDEC=2;       /*调用 means 过程，输出结果保留两位小数*/
var count;                                  /*指定分析变量为 count*/
run;
proc univariate data=chap4.wheat plot ;     /*调用 univariate 过程，输出图形*/
var count;
run;
ods rtf close;                              /*关闭 RTF 文本输出*/
ods listing;                                /*打开列表输出*/
```

MEANS 输出结果如表 4-4 所示。由此可知，有 50 个样本（50 株小麦）进入分析，一株小麦平均大约长有 28（Mean=28.46）粒麦子，最少有 18 粒、最多有 43 粒麦子，小麦单穗粒数标准差为 4.54。

表 4-4　MEANS 过程输出结果

Analysis Variable : count				
N	Mean	Std Dev	Minimum	Maximum
50	28.46	4.54	18.00	43.00

以下分析 UNIVARIATE 过程主要的输出结果：表 4-5 为基本统计量，其中 Location 列下给出了一株小麦上麦粒的均值（Mean）、中位数（Median）和众数（Mode），Variability 列给出相应的标准差（Std Deviation）、方差（Variance）、极差（Range）和上下四分位数差（Interquartile Range）。表 4-6 为分位数值，给出了从最小值（0%，min）到最大值（100%，max）10 个百分位数的值，如 25%分位数为 25，即 25%的小麦单株粒数少于 25。查表可得到与 MEANS 过程得到相同的结论。

表 4-5　基本统计量

Basic Statistical Measures			
Location		Variability	
Mean	28.46000	Std Deviation	4.53652
Median	28.00000	Variance	20.58000
Mode	25.00000	Range	25.00000
		Interquartile Range	7.00000

表 4-6　分位数

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	43
99%	43
95%	36
90%	34
75% Q3	32
50% Median	28
25% Q1	25
10%	23
5%	21
1%	18
0% Min	18

菜单法：

步骤一：选择菜单 Solution|Analysis|Analyst 打开 ANALYST 分析模块。

步骤二：选择菜单 File|Open by SAS name，在弹出的 Select A Member（选择数据集）对话框中单击打开数据集 chap4.Wheat。

步骤三：选择 Statistics|Descriptive|Summary Statistics 命令，弹出如图 4-1 所示对话框。单击变量 count，再单击 Variables 按钮，指定分析变量 count。单击 OK 按钮，则显示结果如表 4-4 所示。

以下介绍图 4-1 所示对话框中的设置选项含义。

（1）单击 Statistics 按钮，则弹出如图 4-2 所示对话框，在此可以通过单击统计量前空白方框的方式选定输出的统计量。选项框中除校正平方和（Corrected sum of squares，简写成 CSS）和未校正平方和（Uncorrected sum of squares，简写为 USS）两个指标外其他的统计量均在 4.1 节中有介绍，请读者自行参考。以下简介 CSS 和 USS，令 x_i（i=1,…,n）代表变量取值，n 为样本量：

CSS：计算公式为 $\mathrm{CSS}=\sum_{i=1}^{n}(x_i-\bar{x})^2$，即为每个值分别减去变量均值以后取平方和。

USS：计算公式为 $USS=\sum_{i=1}^{n}x_i^2$，即为变量取值的平方和。

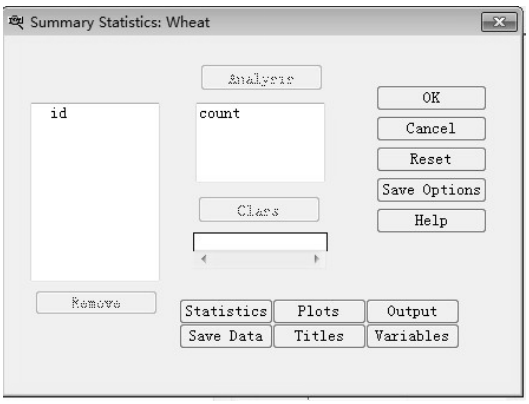


图 4-1　描述性统计分析

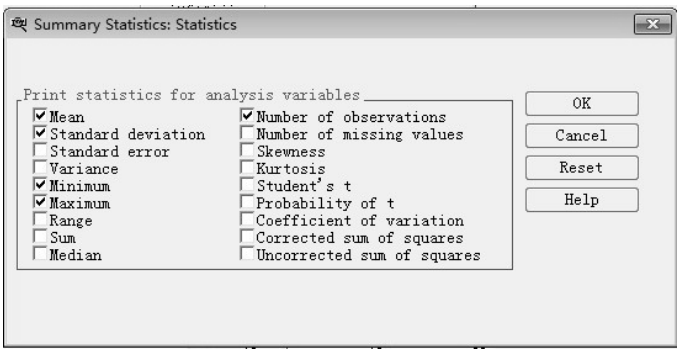


图 4-2　设置输出统计量

在对话框右侧可以进行确定设置（OK 按钮）、取消设置（Cancel 按钮）、重新设置（Reset 按钮）和启动关于此对话框的 SAS 帮助界面（Help 按钮）等操作。这 4 个基本按钮功能以下将不再说明。

（2）单击 Plots 按钮，弹出如图 4-3 所示对话框，在此可以设置输出变量的直方图（Histogram）和触须图（Box-&-whisker plot，俗称箱线图）。

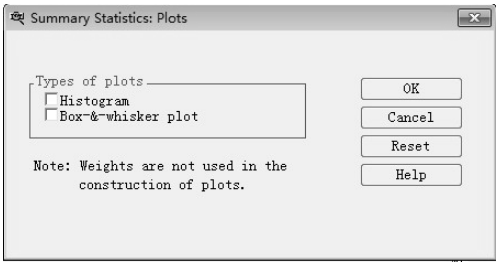


图 4-3　设置图形输出

（3）单击 Output 按钮，弹出如图 4-4 所示对话框，在此可以设置统计量数据的总长度（Field width）和保留的小数点后位数（Number of decimals），以及决定是否打印变量标签（Print variable labels）。

（4）单击 Save Data 按钮，弹出如图 4-5 所示对话框，若选择 Save statistics 后再单击对话框左侧的统计量并单击 Add 按钮将其选中，则分析完毕后将在 work 临时逻辑库中生成一个保存了相应统计量的数据集。

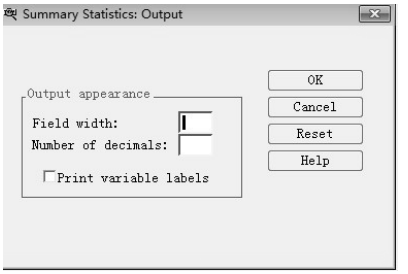


图 4-4　设置统计量输出细节

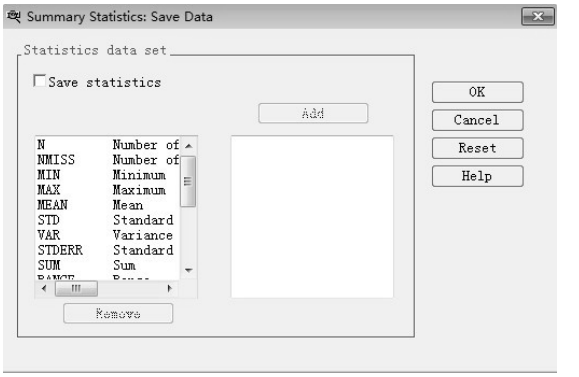


图 4-5　保存统计量到指定数据集

（5）单击 Titles 按钮，将弹出如图 4-6 所示对话框，在此可以设置输出报告的全局标题（Global 选项卡）、描述性统计的标题（Summary Statistics 选项卡）及决定报告中是否包括当前时间、页码和数据筛选细节（Settings 选项卡）。

（6）单击 Variables 按钮，在弹出的对话框（如图 4-7 所示）中可以指定分析的分层变量（BY Group）、加权变量（Weight）和频数变量（Frequency）。

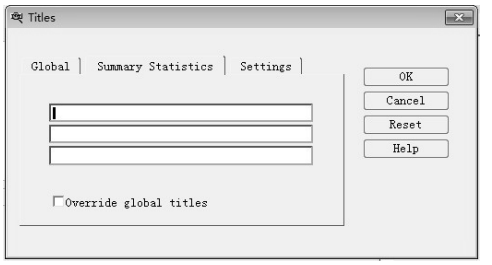


图 4-6　设置标题

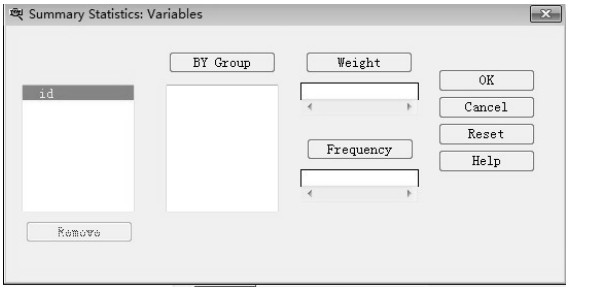


图 4-7　指定分析变量

以上详细介绍了主界面上不同功能按钮对应的对话框的主要设置内容。因 Analyst 分析模块的功能较多，而且不同分析方法的基本设置大同小异，而本书篇幅有限，以下章节再应用到 Analyst 时将不再对对话框一一介绍，希望读者能够举一反三，活学活用。

4.2　描述性统计图形

4.2.1　常见统计图形介绍

除了用统计量指标来描述数据的分布外，我们还可以选择更直观的统计图呈现数据的方式。统计图能用点、线、面、体来形象地表示数量资料信息，方便用户直观地发现和分析问

题。用合适的统计图形来描述数据特征对于分析决策者至关重要，以下为两条统计图选择的原则。

- 资料性质原则：数据资料主要分为定性资料和定量资料，定性资料一般可选用条形图和饼图等；而定量资料一般可选用直方图、折线图、散点图等。
- 图形特征原则：根据图形特征、对照数据特点选择相应绘图方式。表 4-7 总结了几种常见图形的特征供读者参考。

表 4-7　常见图形特征

图 形 类 型	主 要 特 征	举　例
条形图	表示相互独立的统计指标的数量大小。通常纵轴表示数量，横轴为分组标志。绝对数或相对数均可表示数量，图中各长条的高度反映了数值的大小	绘制不同城市的年度生产总值情况
圆饼图	表示事物内部的构成情况。图中每个扇形面积的大小表示百分比数量大小，将 360°圆心角看成是 100%，把每一部分所占的百分比数折算成圆心角的度数，画出对应的扇形	绘制某种食物的不同营养成分所占的百分比
直方图	表示计量资料（测定每个观察单位某项指标值的大小）各组段上的频数的分布情况。图中各长条的面积表示各组数量的大小	绘制某班学生中考平均得分的分布情况
折线图	用于资料中包含两个数量指标，放在横轴上的数量指标通常是时间。纵、横轴上用算术尺度，适用于表示一个或多个事物随着时间的推移在数量上的增减幅度	绘制某城市某年 12 个月的 CPI 情况
散点图	用于资料中包含两个数量指标，且两个变量之间有自变量和因变量之分。通常把自变量放在横轴上，因变量放在纵轴上。将成对的数据点（*X*,*Y*）在 *X* 和 *Y* 直角坐标系中用点表示出来，所以称为散布图或散点图	绘制某一组随机样本的身高和体重的散点图

4.2.2　SAS 过程——GPLOT 过程

使用 GPLOT 过程可生成高分辨率的散点图和折线图等，它的一般使用格式如下：

```
PORC  GPLOT      DATA=数据集;
PLOT      纵坐标变量 Y*横坐标变量 X…</选项列表>;
SYMBOLN   <选项列表>;
AXISN     <选项列表>;
RUN;
```

其中 SYMBOLN 语句用于定义数据点的符号、数据点之间的连接方式及数据点和线的颜色。SYMBOLN 语句的 N 代表 Y*X 两个变量形成数据连线的系列数（取值范围为 1～99，默认值是 1），其后可用的主要选项如表 4-8 所示。

表 4-8　SYMBOL 语句主要选项

选　项	范　例	说　明
V=数据点图形符号	V=PLUS	数据点的符号有 NONE（没有）、PLUS（加号，为系统默认设置）、STAR（星号）、SQUAR（小方块）、DIAMOND（小菱形）、TRANGLE（三角形）、CIRCLE（小圆圈）
I=数据点间连接方式	I=JONE	常用的连接有 NONE（没有）、JOIN（直线）、SPLINE（平滑）、NEEDLE（从数据点到横坐标画垂直线）、RL（直线回归线）、BOX25（盒形线）

续表

选　项	范　例	说　明
W=连线的线宽	*W*=6	设定连线的宽度，单位为像素
H=图形符号的高度	*H*=2.4	设定图形符号的高度，单位为像素
C=定义颜色	C=RED	可定义的颜色有 WHITE、BLACK、BLUE 等

SYMBOL 语句的定义选项数量众多，难以记忆，具体内容可以参见 SAS 在线帮助。进入方式为：单击工具栏上的 Help 图标进入 SAS 帮助系统，选择目录选项卡，SYMBOL 语句具体语法规定的路径为：SAS Products/SAS Graph/Procedure and Statements/SYMBOL。

SYMBOLN 语句一旦被定义就一直保持有效，直到重新定义或退出 SAS 系统。若要取消所有的 SYMBOL 语句定义可提交以下语句：

```
GOPTIONS RESET=SYMBOL;
```

在 GPLOT 绘图过程中，使用 AXISN 语句的选项可以集中、方便地控制和管理坐标轴。常见的使用例子如下：

```
PROC GPLOT      DATA=数据集;
PLOT        Y*X     / VAXIS=AXIS1 HAXIS=AXIS2;
SYMBOLN    <选项列表>;
AXIS1       LABEL=('PRICE') ORDER=(100 TO 1500 BY 200 ) OFFSET=(20,10) ;
AXIS2       LABEL=('DATE') ORDER=('21DEC90'D TO '01JAN93'D BY 98 ) ;
RUN;
```

AXISN 语句的 LABEL 选项规定此轴的标签，ORDER 选项规定此轴的取值范围，OFFSET 选项规定了从原点到此轴的第一个主刻度和从这根轴的最后一个主刻度到最末端的空间大小。例如，语句中的 OFFSET=（20,10），当单位是 PCT 时，表示第一个主刻度空间为图形输出区域的 20%，最后一个刻度的空间为图形输出区域的 10%。

若需要给图形增加第二根纵轴（右轴），程序如下：

```
PORC GPLOT    DATA=数据集;
PLOT     Y1*X     / VAXIS=AXIS1 HAXIS=AXIS2;
PLOT2     Y2*X     / VAXIS=AXIS3;
SYMBOLN    <选项列表>   ;
AXIS1        <选项列表>   ;
AXIS2        <选项列表>   ;
AXIS3        <选项列表>   ;
RUN;
```

注意：当在使用 PROC GPLOT 和 PROC GCHART 等能生成高分辨率图形的过程时，要用到系统的一些特定特征，因此，要求指定图形设备或计算机系统的一些另外可能的选择。用下列 GOPITONS 语句能规定一个图形设备以运行 SAS/GRAPH 软件。

```
GOPTIONS     DEVICE=图形设备;
```

例如，如果我们要指定图形输出设备为 Windows 环境下的彩色打印机：

```
GOPTIONS DEVICE=WINPRTC;
```

选择菜单命令 globals/Options/Global Options 或 globals/Graph/File/Print…/Sas Graph Driver 也可设定图形输出设备。

4.2.3 SAS 过程——GCHART 过程

GCHART 过程用来生成输出在 GRAPH 窗口的高分辨率图形，包括垂直和水平的条形图、饼图、星形图等。GCHART 过程不仅能把输入数据集中的变量值以各种图形的方式显示出来，而且能把输入数据集中的一个或多个变量按分组进行统计计算，再把统计结果以图形的方式显示出来。

GCHART 过程的一般使用格式如下：

```
PROC GCHART      DATA=数据集;
VBAR      变量列表 </<通用选项列表><VBAR 专用选项列表>>;
HBAR      变量列表 </<通用选项列表><HBAR 专用选项列表>>;
BLOCK   变量列表 </<通用选项列表><BLOCK 专用选项列表>>;
PIE       变量列表 </<通用选项列表><PIE 专用选项列表>>;
STAR      变量列表 </<通用选项列表><STAR 专用选项列表>>;
BY       变量列表;
RUN;
```

PROC GCHART 语句后面可以接多个图形要求的语句，如 VBAR、BLOCK 等，即该过程能够对一个指定的数据集画出多种图形。注意：每个图形要求语句后可指定的选项被分成通用选项和专用选项，分别代表可用在每个图形要求语句中的选项和只能在指定图形语句中可用的选项。

通用选项列表中的选项主要包括将要介绍的分组特性选项和变量统计量 TYPE 选项。

专用选项列表中的选项是不同图形语句的特有选项，主要专有选项如表 4-9 所示。

表 4-9 主要专有选项列表

选 项	应 用 语 句	说 明
AXIS=<最小值>最大值	VBAR、HBAR	定义坐标轴的最小值和最大值
GSPACE=间隔数		定义条形组间的间隔空间数大小
ASC/DESC	VBAR、HBAR、PIE	每组内按升序/降序显示条形及有关统计量
G100	VBAR、HBAR、BLOCK	迫使每个组的条状图和统计量加到 100%
BLOCKMAX	BLOCK	定义图中最高块的统计量值
ANGLE=角度数字	PIE、STAR	定义开始逆时针旋转的角度
FILL=SOLID 或 X		每一区域用纯色或交叉线填充
CFILL=颜色		定义图中所有文本的颜色
NOLEGEND	除了 STAR	不输出 SUBGROUP 变量的图例
NOHEADING	BLOCK、PIE、STAR	不输出图表顶部的抬头行

以下三个要素是在应用 GCHART 过程绘图中不可或缺的。

（1）选择图形表示方法：至少选择以下一种图形。

- VBAR 语句——绘制垂直条形图或垂直直方图。
- HBAR 语句——绘制水平条形图或水平直方图。
- BLOCK 语句——绘制块形图。

- PIE 语句——绘制饼图。
- STAR 语句——绘制星形图。

（2）选择变量的统计量：选定图形后，可在语句后面的选项中，通过 TYPE 选项来选择对变量的不同统计量，TYPE 的默认值是 FREQ，然后将统计量以条形或线段表示。统计量的类型有：

- TYPE=FREQ——统计图形变量的各个给定值或间隔的频数。
- TYPE=CFREQ——统计图形变量各个给定值或落入给定区间的累计频数。
- TYPE=PCT——统计图形变量各个给定值或落入给定区间的观测数的百分比。
- TYPE=CPCT——统计图形变量各个给定值或落入给定区间的观测数的累计百分比。
- TYPE=SUM——统计图形变量所有值的总和。
- TYPE=MEAN——统计图形变量所有值的平均值。

例如：VBAR　X　/TYPE=MEAN 语句，指定求 X 变量的平均值，并将其用垂直条形图显示。

（3）选择分组特性：用图形选择语句后的一些选项来控制图形变量的分组：

- DISCRETE——定义一个数字变量为离散变量，则每一个数字值为图形的一个分开的条形或线段。若省略该项，则系统假定变量都是连续的。
- GROUP=变量——对指定变量进行分组。
- SUBGROUP=变量列表——将条形或线段按照指定变量的值分成段。
- SUMVAR=变量——指定用于计算总和或均数的变量。
- MIDPOINTS=数值列表——指定连续性图形变量按数字列表中的中心点数字次序进行排列。在此选项默认情况下，若图形变量是数字型，过程将自动计算各个中心点值；若图形变量是字符型或数字型但选择了 DISCRETE 选项（即离散型数字变量），过程将为每个图形变量的不同值产生一个中心点值。
- LEVELS=数字——指定数字型图形变量的条形或线段个数。

如果没有规定选项 MIDPOINTS=或 LEVELS=，过程自动选择图表的间隔。

4.2.4　SAS 实例——绘制年龄和血压的散点图

散点图（PLOT）又称散布图或相关图，它以散点的分布形式反映变量间的相关关系，根据图中散点的分布形状和密集程度，大致可以判断变量间的协变关系的类型。在回归模型中，常用散点图来描述变量间的相关关系。以下为绘图实例。

例 4-2　根据例 3-10 中的数据绘制年龄和血压的散点图（数据在光盘中的存储路径为“data\chap3\example3_10”）。

编写程序如下所示（在光盘中的存储路径为“proc\chap4\plot”）：

```
proc gplot data=chap3.example3_10;                    /*调用 gplot 绘图过程*/
plot SBP*age/VAXIS=AXIS2 HAXIS=AXIS1;                /*定义绘图变量，指定坐标轴设置*/
SYMBOL   v=star i=none c=black;                       /*指定用黑色星形表示数据，数据间不连接*/
AXIS1      LABEL=('age') ORDER=(20 TO 64 BY 4);       /*定义坐标轴 1 的标签、刻度*/
AXIS2      LABEL=('Systolic blood pressure') ORDER=(100 TO 160 BY 10) ;
RUN;
```

选择菜单 Run|Submit 提交程序，得到如图 4-8 所示散点图。观察发现随着年龄的增长，人

的收缩血压也随之有升高的趋势。

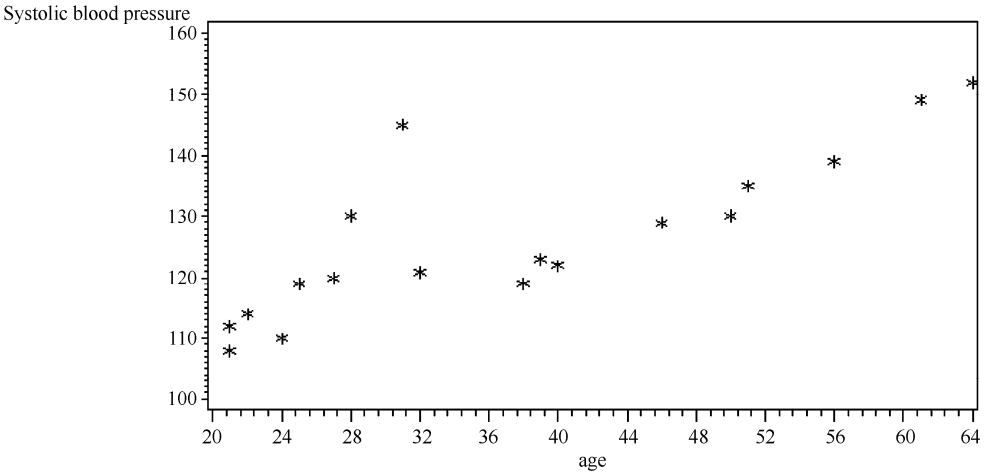


图 4-8　年龄和收缩血压的散点图

4.2.5　SAS 实例——绘制某班学生成绩分布的直方图

直方图（Histogram）又称柱状图、质量分布图，是一种用一系列高度不等的纵向条纹或线段表示数据分布情况的统计图。一般横轴表示数据类型，纵轴表示分布情况。

例 4-3　在一次高考前的英语摸底考试后，某理科实验班 40 名学生的得分情况如表 4-10 所示（相应的 SAS 数据集在光盘中的存储路径为“proc\chap4\histogram”），请据此画出学生得分的直方图，以探索其分布特征。

表 4-10　某班英语摸底考试成绩情况

ID	score	ID	score	ID	score	ID	score	ID	score
01	121	09	124	17	124	25	111	33	107
02	129	10	125	18	124	26	104	34	122
03	107	11	124	19	124	27	126	35	140
04	121	12	138	20	112	28	139	36	117
05	95	13	133	21	129	29	111	37	145
06	113	14	100	22	118	30	106	38	131
07	108	15	103	23	143	31	107	39	124
08	110	16	102	24	110	32	139	40	85

编写 SAS 程序如下所示（其在光盘中的存储路径为“chap4\proc\histogram”）：

```
proc univariate data=chap4.score;   /*调用 univariate 过程*/
var score;                          /*指定分析变量为 score*/
Histogram;                          /*指定绘制分析变量的直方图*/
run;
```

选择 Run|Submit 命令提交程序，则在 Output 结果输出窗口列出变量 score 的描述性统计量的同时将输出直方图，如图 4-9 所示。该图以 12 分为间距，如左侧第二个条状矩形表示有 5%的学生得分在 90～102，而最高的条状矩形表示有 35%的学生得分在 102～114。

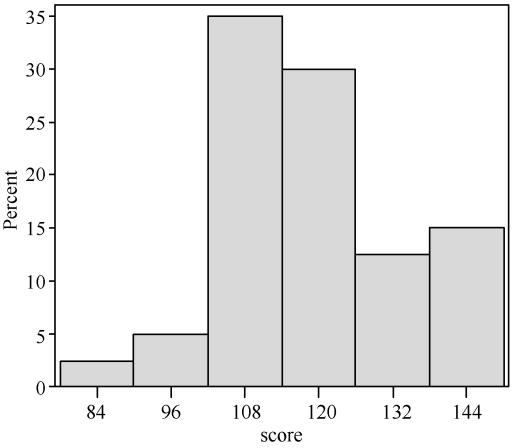


图 4-9　学生成绩分布直方图

4.2.6　SAS 实例——绘制国内生产总值的折线图

折线图又称曲线图，它以线段的升降来说明现象变动情况，主要用于表现在时间上的变化趋势、两个或多个现象之间的依存关系等。

例 4-4　表 4-11 列出了我国自 1978 年改革开放以来至 2010 年的国内生产总值（Gross Domestic Product，GDP），此数据对应的 SAS 数据集在光盘中的存储路径为“data\proc4\line”，请据此绘制折线图。

表 4-11　我国自改革开放以来 GDP 年度数据　　单位：亿元

年度（Year）	GDP	年度（Year）	GDP	年度（Year）	GDP
1978	3645	1989	16992	2000	99215
1979	4063	1990	18668	2001	109655
1980	4546	1991	21781	2002	120333
1981	4892	1992	26923	2003	135823
1982	5323	1993	35334	2004	159878
1983	5963	1994	48198	2005	184937
1984	7208	1995	60794	2006	216314
1985	9016	1996	71177	2007	265810
1986	10275	1997	78973	2008	314045
1987	12059	1998	84402	2009	340507
1988	15043	1999	89677	2010	397983

编写程序如下所示（其在光盘中的存储路径为“proc\chap4\line”）：

```
Proc gplot data=chap4.line;          /*调用 gplot 过程作图*/
Plot GDP*year /haxis=axis1 vaxis=axis2;
/*设定纵轴和横轴变量分别为 GDP、year，且纵轴和横轴的设定参数分别见 axis1、axis2*/
Symbol i=join v=dot l=2 h=0.5;
/*设定数据用直径为 1cm 的点表示，用虚线将数据点连接成折线*/
axis1 label=('年份') order=(1978 to 2010 by 4 );
/*定义 axis1 的标签为“年份”，且取值从 1978—2010 年每隔 4 年取一个刻度*/
axis2 label=('GDP') ;                    /*定义 axis2 的标签为 GDP*/
Title '改革开放以来我国 GDP 走势图';      /*定义标题*/
Run;
```

选择 Run|Submit 命令提交程序，则绘制折线图如图 4-10 所示。该图形直观显示了我国从 1978 年至 2010 年的年度 GDP 的变动趋势，观察到 GDP 随着时间推移呈明显的上升趋势。

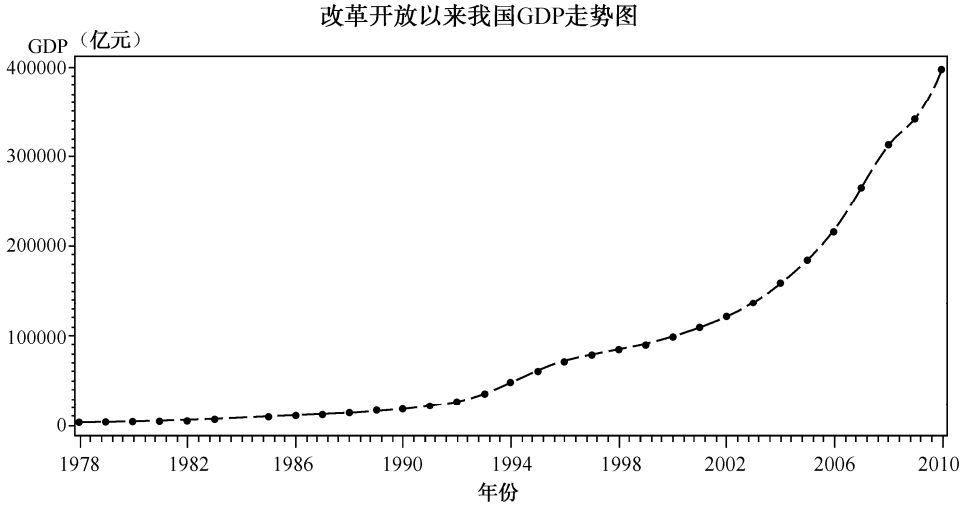


图 4-10　GDP 走势图

4.2.7　SAS 实例——绘制 2009 年 GDP 构成的饼图

饼图的基本特征是以圆的整体面积代表被研究现象的总体，按各个构成部分占总体比重的大小把原面积分割成若干扇形，以表示现象部分对总体的比例关系。

在 SAS 软件中常调用 GCHART 过程中的 PIE 语句绘制饼图，注意指定图形变量 NAME、统计变量 TEST，统计类型为求平均值 MEAN。其他的一些选项主要用于修饰图形，常用的有：

SLICE=——每一块扇形对应的 NAME 值标签方式，=ARROW 表示用一条线指向扇形，=INSIDE 表示标签在扇形内部，=NONE 表示无标签，=OUTSIDE 表示标签在扇形外部。

PERCENT=——每一块扇形相应百分比数的标签方式，有 4 种取值：ARROW、INSIDE、NONE、OUTSIDE。

VALUE=——每一块扇形统计值的标签方式，有 4 种取值：ARROW、INSIDE、NONE、OUTSIDE，取值的含义同 SLINC 语句后的指定选项。

EXPLODE='NAME'——把列出的 NAME 变量值所对应的扇形分离出去，起到强调这一块扇形的作用。注意列表中字符串要与变量 NAME 中值的大小写完全匹配。

ANGLE=——指定第一块扇形的起始角度，默认值为 0。

CTEXT=——设置图中所有文字的颜色。

CFILL=——设置图中所有文本的颜色。

COUTLINE=——设置扇形的轮廓线的颜色。

FILL=SOLID 或 X——设置各个扇形用颜色区别或用交叉线区别。若无此选项，则扇形内为空白。

例 4-5　已知 2009 年的国内生产总值的构成如表 4-12 所示，请绘制呈现不同产业生产总值占国内生产总值的百分比的饼图。

表 4-12　2009 年 GDP 构成　　单位：亿元

国内生产总值	第一产业生产总值	第二产业生产总值	第三产业生产总值
340506.9	35226	157638.8	147642.1

编写程序如下所示（其在光盘中的存储路径为“proc\chap4\pie”）：

```
Proc gchart data=chap4.pie;
Pie type/discrete        /*根据离散型变量 Type 分组绘制饼图*/
    Sumvar=GDP    /*定义计算变量为 GDP*/
    Type=mean     /*饼图的每一个扇形代表的是不同产业生产总值均值，此选项为绘制饼图必选项*/
    Slice=arrow   /*定义用线将扇形和它的标签连接起来*/
    Percent=arrow /*定义用线将扇形和它所代表的百分比连接起来*/
    Value=arrow   /*定义用线将扇形和它所代表的值连接起来*/
    Ctext=black   /*设置图中所有文字为黑色*/
    Cfill=black;  /*设置扇形的轮廓为黑色*/
    title "2009 年 GDP 构成状况";
    Run;
```

选择 Run|Submit 命令提交程序，得到如图 4-11 所示的饼图。

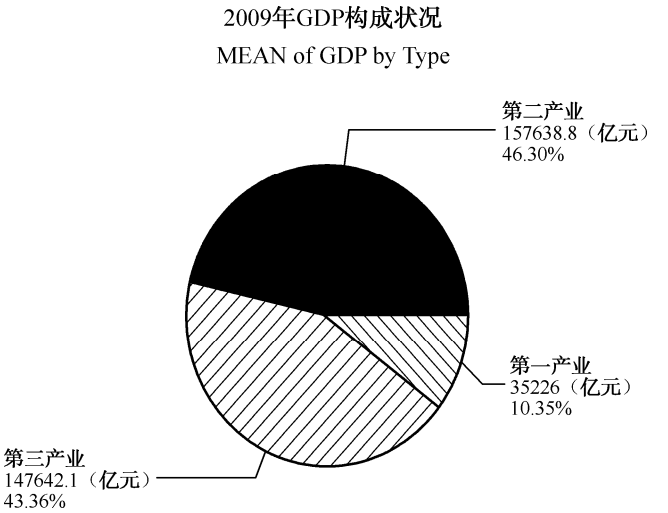


图 4-11　GDP 构成饼图

4.2.8 SAS 实例——绘制某种玉米株高的条形图

条形图利用相同宽度条形的长短或高低表现各个相互对立的统计数据大小或变动情况，可分成水平条形图（又称带形图）和垂直条形图（又称柱形图），分别用 HBAR 和 VBAR 语句实现，条形图主要分成以下三类。

简单条形图——用若干平行、等宽的条状来表示数量对比关系，组间有间隙。

分组条形图——指由每两条或多条组成一组的条形图，组间间隔大，组内条形无间隔或间隔小。绘图过程中使用 GROUP 设置选项。

分段条形图——以条形的全长代表某个变量的整体，条形内部的各分段长短代表各组部分在整体中所占比例，每段用不同颜色或线条表示、之间无间隙，各条之间有间隙。绘图过程中使用 SUBGROUP 设置选项。

例 4-6 从不同地区（A，B，C）抽取不同品种（m，n）的玉米分别 20 株，测量得出玉米的株高（单位为 cm，数据保存路径为“data\chap4\bar”），部分数据如图 4-12 所示。

（1）请绘制不同地区的玉米平均株高的简单条形图。

（2）请分地区绘制不同品种玉米平均株高的分组条形图。

（3）请分品种绘制不同地区的玉米平均株高的分段条形图。

VIEWTABLE: Chap4.Bar

	area	Type	height
1	A	n	155.4
2	A	n	156.7
3	A	n	156.3
4	A	n	153.5
5	A	n	156.6
6	A	n	155.8
7	A	n	155.5
8	A	n	155.4
9	A	n	155.4
10	A	n	154.3

图 4-12 玉米株高部分数据

编写程序如下所示（其在光盘中的存储路径为“proc\chap4\bar”）：

```
Goptions reset=gobal gunit=pct cback=white border
        Htitle=6 htext=3 ftext=swissb colors=(back);    /*预定义图形特征*/
Proc gchart data=chap4.bar;      /*调用 gchart 过程*/
Vbar area/discrete sumvar=height type=mean; /*绘制简单条形图：绘制不同地区玉米平均株高的条形图*/
Run;

Proc gchart data=chap4.bar;
Vbar type/discrete sumvar=height group=area;/*绘制分组条形图：分地区绘制不同品种的玉米株高条形图*/
Run;

Proc gchart data=chap4.bar;
Vbar type/discrete sumbar=height subgroup=area;    /*绘制分段条形图：分品种绘制不同地区的玉米株
                                                    高条形图*/
Run;
```

选择 Run|Submit 命令提交程序，则得到简单条形图（如图 4-13 所示）、分组条形图（如图 4-14 所示）和分段条形图（如图 4-15 所示）。

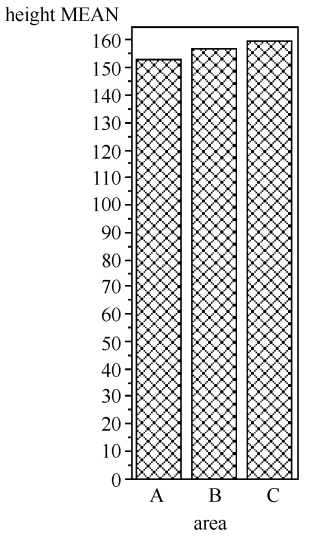


图 4-13　简单条形图

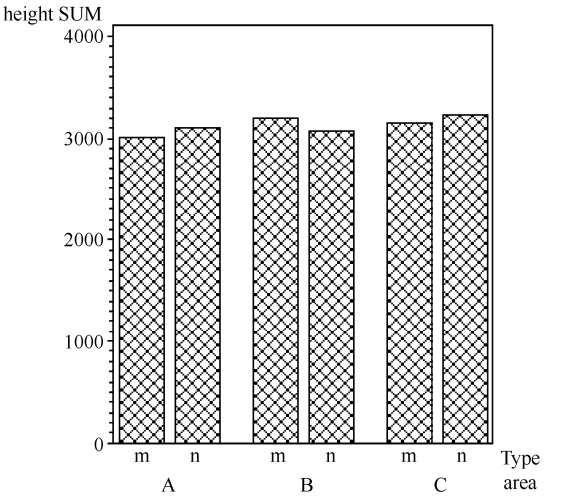


图 4-14　分组条形图

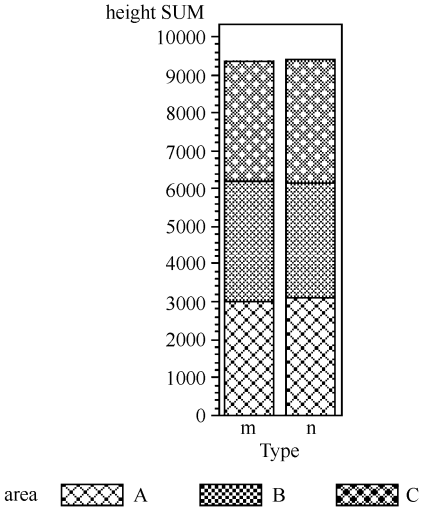


图 4-15　分段条形图

练习题

习题 4-1　某医院的产科统计了 100 名新生儿的体重（单位为 kg），部分数据如表 4-13 所示（包含完整数据的 SAS 数据集在光盘中的存储路径为“data\chap4\weight”）。

（1）请分性别用计算描述性统计指标的方式分析新生儿的体重分布情况；

（2）请分性别绘制新生儿体重的直方图。

表 4-13　新生儿体重部分数据　　单位：kg

编号（ID）	性别（sex）	体重（Weight）	编号（ID）	性别（sex）	体重（Weight）
1	男	1.6	19	男	2.3
2	男	1.6	20	男	2.3
3	男	1.6	21	女	2.4
4	女	1.7	22	女	2.4
5	男	1.8	23	女	2.4
6	男	1.9	24	女	2.4
7	女	2	25	男	2.4
8	女	2	26	女	2.5
9	男	2	27	女	2.5
10	女	2.1	28	女	2.5

（本习题的解答程序在光盘中的存储路径为“proc\chap4\weight”。）

习题 4-2　某调查小组测量了 30 个成年人（18 岁及以上）的两臂展开长度和身高（单位为 cm），数据如表 4-14 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap4\length”）。请据此绘制两臂展开长度和身高的散点图，观察后可初步得到什么结论？

表 4-14　两臂展开长度和身高调查数据　　单位：cm

编号（ID）	身高（Height）	两臂展开长度（Length）	编号（ID）	身高（Height）	两臂展开长度（Length）
1	172.8	175.9	16	164.3	167.3
2	167.8	170.8	17	162.3	165.3
3	166.6	169.6	18	164.0	167.0
4	170.8	173.9	19	161.1	164.0
5	158.8	161.6	20	167.5	170.5
6	175.0	178.2	21	158.7	161.6
7	157.0	159.8	22	161.6	164.5
8	157.4	160.2	23	158.7	161.6
9	153.3	156.0	24	159.3	162.1
10	149.4	152.1	25	166.4	169.4
11	166.4	169.4	26	159.6	162.4
12	161.5	164.4	27	157.2	160.0
13	177.0	180.2	28	146.0	148.6
14	157.8	160.7	29	157.8	160.6
15	157.0	159.8	30	168.6	171.6

（本习题的解答程序在光盘中的存储路径为“proc\chap4\length”。）

习题 4-3　观察某一种植物的生长情况，每隔两天记录它的生长高度，连续记录一个月，所采集的数据如表 4-15 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap4\Growth”）。

试用折线图的方式形象地表现出该植物随时间的生长情况。

表 4-15　某植物生长情况记录数据　　单位：cm

Time（时间点）	Height（高度）	Time（时间点）	Height（高度）	Time（时间点）	Height（高度）
1	3.1	6	5.7	11	11.4
2	3.5	7	6.8	12	12.1
3	3.7	8	7.9	13	12.5
4	4.1	9	8.4	14	12.6
5	4.5	10	9.8	15	13.2

（本习题的解答程序在光盘中的存储路径为“proc\chap4\Growth”。）

习题 4-4　已知某学院的教职工学历构成如表 4-16 所示，请绘制饼图呈现该统计数据结果。

表 4-16　某学院教职工学历构成

学　历	博士	硕士	学士	学士以下
人　数	24	53	42	15

（本习题的解答程序在光盘中的存储路径为“proc\chap4\education”。）

习题 4-5　某课外活动兴趣小组调查了 50 名大三学生一周内课外阅读时间（单位为 h），数据如表 4-17 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap4\time”）。研究目的为比较不同性别、不同专业的学生的阅读时间差异，作为探索性数据分析，请按照以下要求绘制条形图：

（1）请绘制不同性别的学生平均阅读时间的简单条形图；

（2）请分学科绘制不同性别学生平均阅读时间的分组条形图；

（3）请分性别绘制不同学科的学生平均阅读时间的分段条形图。

表 4-17　阅读时间调查数据

编号（ID）	性别（sex）	专业（discipline）	时间（Time）	编号（ID）	性别（sex）	专业（discipline）	时间（Time）
1	女	文科	16	13	女	文科	20
2	女	文科	18	14	女	文科	16
3	女	文科	15	15	女	文科	13
4	女	文科	14	16	女	理科	10
5	女	文科	23	17	女	理科	7
6	女	文科	22	18	女	理科	15
7	女	文科	10	19	女	理科	19
8	女	文科	10	20	女	理科	10
9	女	文科	18	21	女	理科	8
10	女	文科	18	22	女	理科	8
11	女	文科	18	23	女	理科	9
12	女	文科	14	24	女	理科	13

续表

编号（ID）	性别（sex）	专业（discipline）	时间（Time）	编号（ID）	性别（sex）	专业（discipline）	时间（Time）
25	女	理科	12	38	男	文科	14
26	女	理科	21	39	男	文科	11
27	女	理科	14	40	男	文科	11
28	女	理科	15	41	男	理科	12
29	女	理科	13	42	男	理科	11
30	女	理科	13	43	男	理科	7
31	男	文科	13	44	男	理科	6
32	男	文科	14	45	男	理科	6
33	男	文科	8	46	男	理科	7
34	男	文科	15	47	男	理科	4
35	男	文科	13	48	男	理科	6
36	男	文科	9	49	男	理科	9
37	男	文科	4	50	男	理科	11

注：时间单位为 h。

（本习题的解答程序在光盘中的存储路径为“proc\chap4\time”。）

第 5 章　参数估计与假设检验

在统计学研究中，抽样调查的目的是根据样本推断总体的信息，而统计推断的最主要方式为参数估计和假设检验。参数估计指根据从总体中抽取的样本来估计总体分布中包含的未知参数，或者是拟合特定的模型并估计模型中包含的系数，它分为点估计和区间估计。点估计主要有矩估计和极大似然估计。区间估计即按一定的概率估计总体参数在哪个范围内。假设检验是统计推断的重要组成部分，它被分为参数假设检验和非参数假设检验。参数假设检验是对总体分布函数中的未知参数提出某种假设，然后利用样本提供的信息对所提出的假设进行检验，根据检验结果做出接受或拒绝原假设的判断。非参数假设检验主要是对总体分布函数形式或总体的性质提出某种假设。假设检验的一般步骤为：

（1）提出原假设和备择假设。

（2）确定适当的检验统计量并计算它的值。

（3）规定显著性水平。

（4）做出统计决策。

本章介绍基本的参数估计和假设检验，重点介绍单样本均值和方差的区间估计、*t* 检验（包括单样本 *t* 检验、独立样本 *t* 检验和配对样本 *t* 检验）和正态分布拟合检验。首先简介基本的数理统计模型，再与实例结合起来以编程和菜单操作的方式实现上述分析。

5.1　TTEST 过程

SAS 系统中很多统计过程都包含有参数估计和假设检验。本章主要应用第 4 章介绍的 MEANS 和 UNIVARIATE 过程，以及以下的 TTEST 过程。

TTEST（*t* 检验）过程的一般语法格式为：

```
PROC TTEST   DATA=数据集 <选项列表> ;
CLASS      变量列表;
VAR        变量列表;
BY         变量列表;
RUN;
```

PROC TTEST 语句后的主要控制选项如表 5-1 所示。

表 5-1　TTEST 后的主要控制选项

选　项	意　义
COCHRAN	在方差不等的情况下用 COCHRAN 和 COX 方法计算近似的 t' 统计量的近似概率水平
SIDES=	取值“2”代表双侧检验、“L”代表左侧检验、“U”代表右侧检验
DIST=NORMAL\LOGNORMAL	分布检验，若 DIST=NORMAL，则为正态检验；若取为 DIST= LOGNORMAL，则为对数正态检验

$H_1:\bar{\xi}<\mu$（样本均值小于总体均值）等此类备择假设取小于号（或小于等于号）的在本书中统称为左侧假设，而 $H_1:\bar{\xi}>=\mu$（样本均值大于等于总体均值）等此类备择假设取大于号（或大于等于号）的假设类型在本书中统称为右侧假设。

TTEST 过程中使用的语句意义解释如下：

CLASS 语句——定义分组变量（字符型或数值型变量），它有且只有两个水平。

BY 语句——得到由 BY 变量定义的几个观察组，分别分析。

VAR 语句——定义分析变量。在此语句默认时系统将分析输入数据集中所有数值型变量（除了 CLASS 语句中已定义的）。

5.2 基本的参数区间估计

本节介绍总体参数区间估计，即按一定的概率估计总体的参数取值范围，这个范围称为置信区间，这个概率称为可信度或置信度，用 $1-\alpha$ 表示。比较常见的是求取 95%（$\alpha=0.05$）或 99%（$\alpha=0.01$）置信区间。以下介绍总体均值和方差的区间估计。

5.2.1 总体均值的区间估计

总体均值用 μ 表示，其点估计为样本均值：$\hat{\mu}=\bar{x}$。总体均值的区间估计由已知条件不同而采取不同的方法，主要分成以下两种情况。

1．总体服从正态分布

若总体服从正态分布 $N(\mu,\sigma^2)$ 且总体方差 σ^2 已知，则样本的均值分布为：

$$\bar{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$$

对 $\bar{X}$ 变量进行标准化处理，得到 U 检验统计量：

$$U=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$

于是总体均值的 $1-\alpha$ 置信区间为：

$$\left(\bar{X}-u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\bar{X}+u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

若总体服从正态分布 $N(\mu,\sigma^2)$ 但总体的方差 σ^2 未知，这是实际中最常见的情形，此时可用样本标准差 $S\left(S=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$ 来代替总体标准差 σ，由抽样分布得到 t 检验统计量：

$$t=\frac{\bar{X}-\mu}{S/\sqrt{n}}\sim t(n-1)$$

于是总体均值的 $1-\alpha$ 置信区间为：

$$\left(\bar{X}-t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}},\bar{X}+t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}\right)$$

2．总体不服从正态分布

实际上，大多数情形下总体并不服从或仅近似服从正态分布。此时根据中心极限定理，只要样本容量n足够大，样本均值$\bar{X}$的抽样分布就近似服从正态分布。若方差σ^2已知（根据历史资料或经验得到），则可用公式：

$$\left(\bar{X}-u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\bar{X}+u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

来计算总体均值的$1-\alpha$置信区间。若σ^2未知，则用样本标准差S来代替总体标准差σ，总体均值的$1-\alpha$置信区间为：

$$\left(\bar{X}-u_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}},\bar{X}+u_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right)$$

5.2.2 总体方差的区间估计

若总体服从正态分布$N(\mu,\sigma^2)$，由于S^2是σ^2的最优无偏估计量，而根据抽样分布定理有$\frac{(n-1)S^2}{\sigma^2}\sim\lambda^2(n-1)$，若需要求$\sigma^2$的$1-\alpha$置信区间，即有：

$$P\left\{\lambda^2_{1-\alpha/2}(n-1)<\frac{(n-1)S^2}{\sigma^2}<\lambda^2_{\alpha/2}(n-1)\right\}=1-\alpha$$

可写作：

$$P\left\{\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}<\sigma^2<\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right\}=1-\alpha$$

于是方差σ^2的置信度为$1-\alpha$的置信区间为：

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)},\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right)$$

若总体的均值已知，则将$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2$代入公式中。

若总体的均值未知，则将$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$代入公式中，此为最常见的情况。

5.2.3 SAS 实例——求均值和方差的 95%置信区间

分析者时常需要根据样本数据来对总体的聚集和离散情况做一个较为精确的估计，如零件质检员希望通过随机抽取一部分样本来得到此零件的平均长度及变化幅度；市场调查人员希望通过科学的抽样方法得到某种商品的价格，以此估计此商品在整个市场的价格集中和离散程度等。这类问题归结为单正态总体的均值和方差（标准差）的区间估计。

例 5-1 某市场调研机构为了调查某一品牌的 250mL 食用油的市场价格情况，随机抽取了 10 所农贸市场中的 40 家粮油店，记录下此食用油的售价（变量：price）如表 5-2 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap5\oil”）。请估计该食用油市场售价均值和方差的 95%置信区间。

表 5-2 某品牌食用油价格抽查情况 单位：元

68	69	68	65	66	69	68	67	69	65
70	66	69	68	67	68	66	66	67	71
68	68	69	69	70	66	66	66	68	67
66	65	69	69	64	67	68	67	68	68

编程法：

编写以下程序（其在光盘中的存储路径为“proc\chap5\oil”）：

```
proc ttest data=chap5.oil;          /*调用 ttest 过程*/
var price;                          /*定义分析变量为 price*/
run;

ods select BasicIntervals;
proc univariate data=chap5.oil cibasic(alpha=.05);
var price;
run;
```

选择 Run|Submit 命令提交程序，将以此输出变量 price 的简单描述性统计量，均值、标准差的置信区间（如表 5-3 所示）和零均值的 t 检验结果（假设检验将于下节介绍）。观察可知该品牌 250mL 食用油价格的 95%置信区间为（67.0，68.0）、标准差的 95%置信区间为（1.2，2.0）（结合本例的实际背景，结果仅保留一位小数）。换言之，调查员有 95%的把握认为该食用油的均价在 67～68 元，而且价格的变动幅度在 1.2～2 元。

表 5-3 均值和标准差的 95%置信区间 单位：元

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
67.5000	66.9982	68.0018	1.5689	1.2852	2.0146

注意：系统默认的置信水平 α 为 0.05，用户也可自行设置，若将本程序的第一行改为“Proc ttest data=chap5.oil alpha=0.01;”，则程序运行成功后系统将输出正态总体的均值和标准差的 99%置信区间。

与此同时，UNIVARIATE 过程也会给出一致的结果。

菜单法：

步骤一：选择菜单 Solutions|Analysis|Analyst 命令进入 Analyst 分析模块。选择 File|Open By SAS name|chap4|oil 命令，打开数据集 chap4.oil。

步骤二：求解一个正态总体均值的置信区间。

选择 Statistics|Hypothesis Test|One-Sample t-test for Mean 命令，弹出如图 5-1 所示对话框，单击选中 price 为分析变量。Hypotheses 为假设检验选项，将在下节介绍。

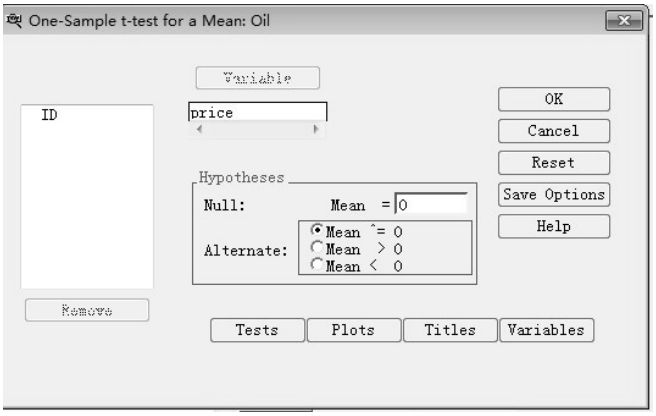


图 5-1　单样本 *t* 检验

单击 Tests（检验）按钮，弹出如图 5-2 所示对话框，选择置信区间（Confidence intervals）选项框中的 Interval（双侧），即设置计算双侧置信区间，本选项框其他选项意义为：None（无）——不计算置信区间、Lower bound（下限）——计算置信下限、Upper bound（上限）——计算置信上限。置信度（Confidence level）选项采用系统默认的 95%。单击 OK 按钮保存设置并返回如图 5-1 所示对话框。单击 OK 按钮提交设置得到分析变量 price 的描述性统计量与均值的 95%置信区间。与编程法得到的结果一致。

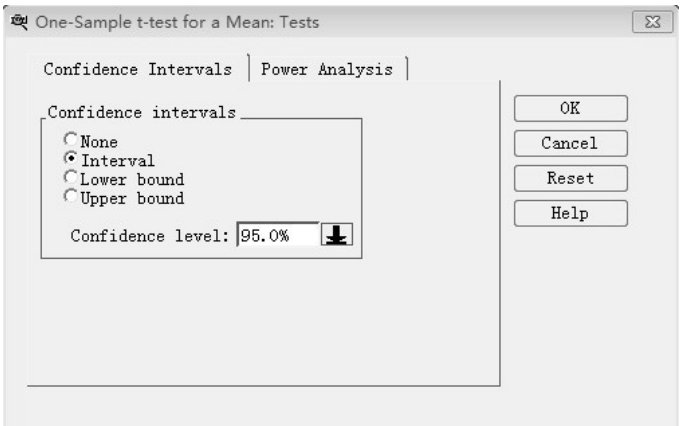


图 5-2　设置置信限和置信水平

步骤三：求解一个正态总体方差的置信区间。

选择 Statistics|Hypothesis Test|One-Sample test for a Variance 命令，弹出如图 5-3 所示对话框，单击选中变量 price 为分析变量，在 Hypotheses（假设检验）选项框中任意填入 1（注：此项必填，若空白将不能完成操作，而且必须输入非零数值），单击 Intervals（区间）按钮，弹出如图 5-2 所示对话框，类似的，在此可以设置计算置信区间的类型及置信水平，设置计算 95%置信区间后，单击 OK 按钮保存设置，即采用系统默认的求双侧的 95%置信区间，返回图 5-3 所示对话框，单击 OK 按钮提交设置，则输出结果简单描述性统计量、假设检验的结果及总体方差的 95%置信区间，注意由于本例的目的为求标准差的 95%置信区间，因此将方差的置信下限和置信上限分别开平方，即得到标准差的置信区间。

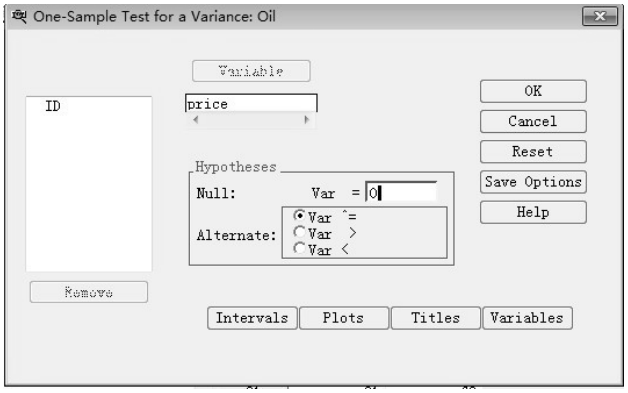


图 5-3　单样本方差检验

5.3　基本假设检验

本节介绍常见假设检验——均值检验、方差齐性检验、分布检验。SAS 系统中主要由 TTEST 过程来实现单样本和两样本均值检验和方差齐性检验，可调用 UNIVARIATE 过程实现正态分布检验。以下分别介绍 t 检验、方差齐性检验和正态分布检验的基本原理。

5.3.1　t 检验

1．单样本 t 检验

为检验某正态总体均值是否等于（或大于、小于）某特定值 μ，从总体中抽取部分样本，由于样本均值 $\bar{x}$ 为总体均值的无偏估计，则转换成检验 $\bar{x}$ 是否等于（或大于、小于）μ，以下给出检验 $\bar{x}$ 是否等于 μ 的推导过程。

原假设 $H_0:\bar{x}=\mu$，备择假设 $H_1:\bar{x}\neq\mu$

若令 $d_i=x_i-\mu$，则 $\bar{d}=\bar{x}-\mu$。则该假设检验可写作：

原假设 $H_0:\bar{d}=0$，备择假设 $H_1:\bar{d}\neq 0$

由抽样分布定理得到 t 检验统计量：

$$t_{检}=\frac{\bar{d}}{S_d/\sqrt{n}}\sim t(n-1)$$

其中，

$$S_d=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i-\bar{d})^2}$$

再根据 $t_{检}$ 求出对应的 $P(|t|\geqslant t_{检})$，将 $P(|t|\geqslant t_{检})$ 与设定的显著性水平 α（一般取 0.01 或 0.05）进行比较，若 $P(|t|\geqslant t_{检})<\alpha$，则拒绝原假设，认为该样本均值不等于特定值；若 $P(|t|\geqslant t_{检})\geqslant\alpha$，则接受原假设，认为该样本均值等于特定值。

当检验原假设 $H_0:\bar{x}=\mu$ 或 $H_0:\bar{x}\leqslant\mu$，备择假设 $H_1:\bar{x}>\mu$ 时，经过类似推导得到 $P(t\geqslant t_{检})$，

将其与α比较得出结论；当检验原假设$H_0:\bar{x}=\mu$或$H_0:\bar{x}\geqslant\mu$，备择假设$H_1:\bar{x}<\mu$，计算得出$P(t\leqslant t_{检})$，将其与α比较得出结论。

2．配对样本 *t* 检验

该检验适用于配对试验设计，主要为两种情形：一是按一些非试验因素条件将受试对象配对，给予每对中的个体以不同的处理，如按照受试者的年龄、身高、体重等指标进行配对试验比较两种处理（x_{1i}和x_{2i}）的效果；二是自身对照，观察同一受试对象（指标）不同时间、试验前后的变化，如自身对照中比较试验前后某指标（x_{1i}和x_{2i}）的变化，首先求出数据之差$d_i=x_{1i}-x_{2i}$，则后续的假设检验分析过程参见以上单样本 *t* 检验推导过程。

3．独立样本 *t* 检验

该检验方法适用于完全随机化设计，设总体$X_1\sim N(\mu_1,\sigma_1^2)$，$X_2\sim N(\mu_2,\sigma_2^2)$，从两个总体中分别抽取一定量的样本，根据样本信息检验两总体均值是否相等（或某样本均值是否大于或小于另一样本），以下推导检验两样总体均值是否相等的过程。

原假设$H_0:\mu_1=\mu_2$，备择假设$H_1:\mu_1\neq\mu_2$。

如果σ_1^2和σ_2^2都已知，则

$$\bar{X}_1-\bar{X}_2\sim N\left(\mu_1-\mu_2,\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)$$

经标准化变换：

$$U_{检}=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}\sim N(0,1)$$

再根据$U_{检}$求出对应的$P(|U|\geqslant U_{检})$，将$P(|U|\geqslant U_{检})$与设定的显著性水平α进行比较，若$P(|U|\geqslant U_{检})<\alpha$，则拒绝原假设，认为两个总体均值相等；若$P(|U|\geqslant U_{检})\geqslant\alpha$，则接受原假设，认为两个总体均值不相等。

如果σ_1^2和σ_2^2都未知且等值，即$\sigma_1^2=\sigma_2^2=\sigma^2$，$\sigma$是需要估计的未知值。由于$s_1^2$和$s_2^2$都是$\sigma^2$的无偏估计，都包含$\sigma^2$的信息，则使用合并方差估计法：

$$s_p^2=\frac{\sum(x_1-\bar{x}_1)^2+\sum(x_2-\bar{x}_2)^2}{n_1+n_2-2}=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

然后用样本合并标准差s_p来代替U中的总体标准差σ，得到统计量：

$$t=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{s_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}\sim t(n_1+n_2-2)$$

如果σ_1^2和σ_2^2都未知且方差不等$\sigma_1^2\neq\sigma_2^2$。用s_1^2和s_2^2分别估计σ_1^2和σ_2^2后

$$t^*=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}}\sim t(l)$$

这时 t^* 就不再服从 $N(0,1)$ 分布了，但形式接近 t 统计量，可用以下修正后的 t^* 统计量做出合适的统计推断：

$$t^* = \frac{\frac{s_1^2}{n_1}t_\alpha(n_1-1)+\frac{s_2^2}{n_2}t_\alpha(n_2-1)}{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \sim t(n_1+n_2-2)$$

另外，Satterthwaite 设法用 t 统计量去拟合，发现若取

$$l = \left(\frac{s_2^2}{n_1}+\frac{s_2^2}{n_2}\right)^2 / \left(\frac{s_1^4}{n_1^2(n_1-1)}+\frac{s_2^4}{n_2^2(n_2-1)}\right)$$

当 l 的计算结果为非整数时取最接近的整数，则 t^* 近似服从自由度为 l 的 t 分布。当样本数 n_1 和 n_2 较大时，式中的 l 值也将随之增大，当 $l \geqslant 30$ 时，自由度为 l 的 t 分布就很接近于正态分布 $N(0,1)$，故在 n_1 和 n_2 较大时，我们将认为 t^* 统计量服从 $N(0,1)$ 分布。

得到统计量以后再求相应的检验 P 值，将其和显著性水平进行比较得出检验结论。

5.3.2 两样本方差齐性检验

两样本均值 t 检验的前提之一是两样本所代表的正态总体方差相等，因此在采取 t 检验时应事先检验两个方差是否相等，称为方差的齐性检验。

已知两独立的正态总体，设为正态分布总体 $X_1 \sim N(\mu_1,\sigma_1^2)$，$n_1$ 个样本均值为 $\overline{X}_1$、样本方差为 s_1^2；另一正态分布总体 $X_2 \sim N(\mu_2,\sigma_2^2)$，$n_2$ 个样本均值为 $\overline{X}_2$、样本方差为 s_2^2。假定 μ_1 和 μ_2 未知，检验的原假设 $H_0:\sigma_1^2=\sigma_2^2$，备择假设 $H_1:\sigma_1^2 \neq \sigma_1^2$。由于：

$$\frac{(n_1-1)}{\sigma_1^2}s_1^2 = \sum_i^{n_1}\left(\frac{x_{1i}-\overline{x}_1}{\sigma_1}\right)^2 \sim \chi^2(n_1-1)，\ \frac{(n_2-1)}{\sigma_2^2}s_2^2 = \sum_i^{n_2}\left(\frac{x_{2i}-\overline{x}_2}{\sigma_2}\right)^2 \sim \chi^2(n_2-1)$$

构造统计量：

$$F_{检} = \frac{s_1^2}{s_2^2}\cdot\frac{\sigma_2^2}{\sigma_1^2} \sim F(n_1-1,n_2-1)$$

在原假设 $H_0:\sigma_1^2=\sigma_2^2$ 为真的情况下 $F_{检} = \frac{s_1^2}{s_2^2} \sim F(n_1-1,n_2-1)$，对于给定显著水平 α，若 $F_{检} \leqslant F_{\frac{\alpha}{2}}(n_1-1,n_2-1)$ 或 $F_{检} \geqslant F_{1-\frac{\alpha}{2}}(n_1-1,n_2-1)$ 则拒绝原假设。

不拒绝 H_0 时，满足 t 检验的前提条件方差齐性的条件，则计算的 t 统计量及统计推断可靠；拒绝 H_0 时，认为两个总体方差不齐，此时不能直接进行 t 检验，而应该采取适当的措施，如检查试验的本身、进行数据转换或用非参数统计分析方法等。

5.3.3 正态分布检验

该检验用于判断总体分布是否为正态分布。在 SAS 系统中，小样本（SAS 默认样本量小于 2000）情形时推荐使用 Shapiro-Wilk 的 W 检验，而大样本（SAS 默认样本量大于等于 2000）情

形时有基于经验分布函数（Empirical Distribution Function，EDF）的 Kolmogorov-Smirnov 检验、Anderson-Darling 检验和 Cramér-von Mises 检验。

原假设 H_0：总体 X 服从正态分布，备择假设 H_1：总体 X 不服从正态分布。

W 检验：在抽取小样本时，Shapiro 和 Wilk 提出用如下的 W 统计量：

$$W=\frac{\left[\sum_{i=1}^{n}(a_i-\overline{a})(x_i-\overline{x})\right]^2}{\sum_{i=1}^{n}(a_i-\overline{a})^2\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

系数 a_i 按标准正态分布构造，均值为 0，标准差为 1，且是对称值。W 统计量可以看成是数对（a_i, x_i）相关系数的平方，它的取值在 0～1。在 H_0 原假设为真时，W 的取值应接近于 1，并根据统计量分布得出显著性检验 P 值。

以下简单介绍 Kolmogorov-Smirnov 检验的基本原理，要检验样本是否来自于某个已知分布 $F_0(x)$，用 $S(x)$ 代表该组数据的经验分布，一般来说随机变量 $X_1, X_2, \cdots, X_n$ 的经验分布函数定义为 $S(x)=\dfrac{X_i \leqslant x\text{的个数}}{n}$，于是 Kolmogorov-Smirnov 统计量可以表示为：$D=\sup|S(x)-F_0(x)|$，然后再由大样本情形下，有渐进分布：

$$P(\sqrt{n}D_n<x)\sim K(x)$$

而分布函数 $K(x)=\begin{cases}0 & x<0\\ \sum_{j=-\infty}^{\infty}(-1)^j\exp(-2j^2x^2) & x>0\end{cases}$

再据此计算出显著性检验 P 值，将其和 α 进行比较得到最终的检验结果。

5.3.4　SAS 实例——检验水稻单株产量是否为特定值

例 5-2　从一亩试验田中随机抽取 50 株水稻，测出其单株产量（数据如表 5-4 所示，相应的 SAS 数据集在光盘中的存储路径为“data\chap5\rice”）。请问该水稻单株产量是否等于 250g？（α =0.05）

表 5-4　水稻单株产量　　单位：g

245.3	257.2	256.4	257.7	257.8	265.3	265.3	251.6	257.8	250.9
252.5	262.2	253.3	255.3	245.1	260.1	253.4	259.1	254.1	250.0
255.4	253.7	248.1	256.6	254.6	259.4	257.9	256.7	252.3	244.3
250.8	257.7	257.7	256.4	253.4	255.1	256.5	249.9	257.7	258.2
257.6	259.9	259.7	248.4	252.4	254.8	256.8	249.2	261.6	255.7

解析：本实例为典型的单样本 t 检验情形。原假设为：该水稻单株产量=250g，备择假设为：该水稻单株产量≠250g。以下将调用 TTEST 过程以编程的方式及调用 Analyst 模块以菜单操作的方式进行分析。

编程法：

编写如下程序（其在光盘中的存储路径为“proc\chap5\rice”）：

```
data meantest;                    /*新建临时数据集 meantest*/
set chap5.rice;                   /*导入数据集 chap5.rice*/
dif=weight-250;                   /*新建变量 dif 为变量 weight 的值减去 250*/
run;

/*方法一：调用 means 过程完成单样本 t 检验*/
proc means data=meantest t prt;   /*指定输出零均值检验的 t 统计量和 P 值*/
var dif;                          /*指定分析变量为 dif*/
run;

/*方法二：调用 ttest 过程完成单样本 t 检验*/
proc ttest data=meantest;         /*调用 ttest 分析过程*/
var dif;                          /*指定分析变量为 weight*/
run;
```

以上分别调用了 MEANS 过程和 TTEST 过程进行单样本 *t* 检验，这两者的主要区别在于：MEANS 过程只能进行双侧检验，而 TTEST 过程在 proc ttest 语句后指定 side=L/U/2 来进行左侧、右侧和双侧检验。

选择 Run|Submit 命令提交程序。不同的分析过程得到了相同的检验结果，*t* 统计量为 7.89，对应的 *P* 值小于 0.0001，明显小于显著性水平 0.01，则拒绝原假设，认为样本均值在 0.01 的显著性水平下不等于 250g，即认为该水稻的单株产量不等于 250g。

菜单法：

步骤一：选择 Solutions|Analysis|Analyst 命令进入 Analyst 操作界面。

步骤二：选择 File|Open By SAS name 命令，单击打开数据集 chap5.rice。

步骤三：选择 Statistics|Hypothesis Test|One Simple t-test for mean 命令，弹出如图 5-4 所示对话框，单击选中变量 weight 后单击分析变量（Variables）按钮。在 Hypotheses（假设检验）选项框下 Null（原假设）后面填入 250，根据题意在 Alternate（备择假设）选择 Mean=250。单击 OK 按钮则显示结果如图 5-5 所示：样本的部分描述性统计量（样本量、均值、标准差、标准误），原假设和备择假设，*t* 统计量、自由度和 *P*（Prob>*t*）值。由于 *P* 值（<0.0001）明显小于显著性水平 0.01，则与编程方法得到的结果一致。

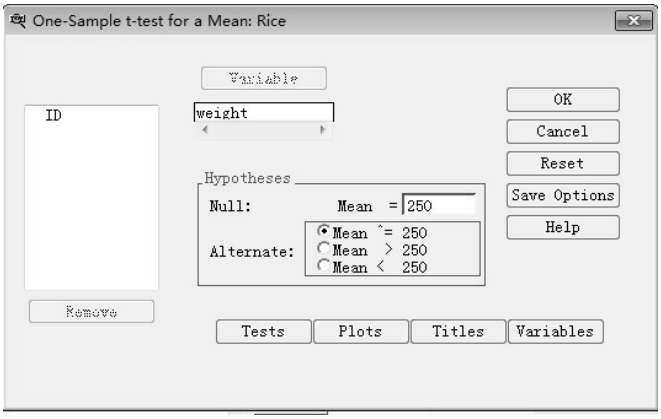


图 5-4　单样本 *t* 检验对话框

Sample Statistics for weight

N	Mean	Std. Dev.	Std. Error
50	255.18	4.66	0.66

Hypothesis Test

Null hypothesis:　Mean of weight = 250
Alternative:　Mean of weight ^= 250

t Statistic	Df	Prob > t
7.856	49	<.0001

图 5-5　检验结果

5.3.5　SAS 实例——比较不同方法的减肥效果

例 5-3　为了比较不同运动方式减肥的效果，某健身中心将需要运动减肥的 40 名会员随机分成了两组，每组 20 人，在一个月以内分别采取每天慢跑一小时（方式 A）和每天游泳一小时（方式 B）两种措施，并记录下他们运动后的减重情况（减重单位为 kg，数据如表 5-5 所示，相应的 SAS 数据集在光盘中的存储路径为“data\chap5\loseweight”）。请问这两种运动减肥方法的效果是否存在显著差异？（α =0.05）

表 5-5　运动减重记录　　单位：kg

运动方式	减　重	运动方式	减　重	运动方式	减　重	运动方式	减　重
A	0.9	A	2.1	B	3.4	B	2.0
A	4.0	A	1.8	B	3.6	B	2.0
A	1.2	A	0.7	B	2.0	B	1.0
A	1.4	A	1.5	B	1.8	B	3.3
A	1.0	A	2.5	B	2.1	B	2.2
A	2.0	A	2.2	B	3.8	B	2.0
A	1.6	A	2.4	B	2.3	B	2.1
A	1.0	A	3.2	B	3.5	B	1.8
A	3.5	A	1.7	B	2.2	B	2.7
A	0.8	A	2.6	B	2.6	B	2.9

解析：本实例中，将训练者随机分成了不相关的两组，再来探索采用了不同的减重运动方式的两组队员减重情况是否有显著差异，因此应该采用独立样本 *t* 检验。

编程法：

编写程序如下所示（其在光盘中的存储路径为“proc\chap5\loseweight”）：

```
Proc ttest data=chap5.loseweight;        /*调用 ttest 过程*/
Class type;                              /*定义分类变量为 type*/
Var lose_weight;                         /*指定分析变量为 ram*/
Run;
```

选择 Run|Submit 命令提交程序，结果如表 5-6 至表 5-9 所示。系统首先按变量 type 的取值

分类计算出分析变量 loseweight 的简单描述性统计量（表 5-6：从左到右依次为样本量、均值、标准差、标准误、最小值和最大值），type 变量下的 Diff（1−2）行为两组观测的均值差的描述性统计量（均值差、均值差的标准差和标准误）。

表 5-6　描述性统计量

type	N	Mean	Std Dev	Std Err	Minimum	Maximum
A	20	1.9050	0.9237	0.2065	0.7000	4.0000
B	20	2.4650	0.7365	0.1647	1.0000	3.8000
Diff (1−2)		−0.5600	0.8353	0.2642		

表 5-7 为变量均值和标准差的 95%置信区间（分类结果）。最后两行为分别用 Pooled 方法和 Satterthwaite 方法计算的两样本（即 A 组和 B 组样本）均值差的 95%置信区间，若两样本方差之间差异不显著，则选用 Pooled 方法计算的结果；若存在显著差异，则选用 Satterthwaite 方法计算的结果。

表 5-7　均值与方差的 95%置信区间

type	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
A		1.9050	1.4727	2.3373	0.9237	0.7024	1.3491
B		2.4650	2.1203	2.8097	0.7365	0.5601	1.0757
Diff (1−2)	Pooled	−0.5600	−1.0947	−0.0253	0.8353	0.6827	1.0765
Diff (1−2)	Satterthwaite	−0.5600	−1.0956	−0.0244			

虽然表 5-8 先于表 5-9 输出，在实际应用中，应该首先看两样本的方差齐性 *F* 检验结果（如表 5-9 所示），*F* 检验的原假设为“A 和 B 样本的方差相等”，对应的检验 *P* 值为 0.332，大于设定的显著性水平 0.05，则接受原假设，认为两组观测的方差差异不显著。因此选择 Pooled（合并方差）方式进行 *t* 检验（如表 5-8 所示第二行），对应的双边检验 *P* 值为 0.0409，因此在显著性水平 0.05 时，这两种运动方式的减重效果差异显著。

表 5-8　*t* 检验结果

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	38	−2.12	0.0406
Satterthwaite	Unequal	36.205	−2.12	0.0409

表 5-9　方差齐性检验

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	19	19	1.57	0.3320

菜单法：

步骤一：选择菜单 Solutions|Analysis|Analyst 命令进入 Analyst 操作界面。

步骤二：选择 File|Open By SAS name 命令，单击打开数据集 chap5.loseweight。

步骤三：选择 Statistics|Hypothesis Test|Two Simple t-test for mean 命令，弹出如图 5-6 所示对话框。在此对话框中的 Groups are in（比较组所在）选项下可以设置进行 *t* 检验的两组的存储的变量个数（若设置 1 个则需要设置一个分类变量；若设置为 2 个，则需要将两组值分别存储到两个变量中，如将 A 组和 B 组队员减重情况分别存储在两个不同变量中），采用系统默认的 One variable（单变量）。单击选定变量 lose_weight 为分析变量、变量 type 设定为分组变量。在 Hypotheses（假设检验）选项框中的 Mean1−Mean2（均值差）选项框内填入 0（用户可根据实际情况自主设定）；Alternative（备择假设）后面有三种选择：均值差不等于零、大于零和小于零，采用系统默认的均值差不等于零。单击 OK 按钮提交设置则将显示与编程方法一致的结果。

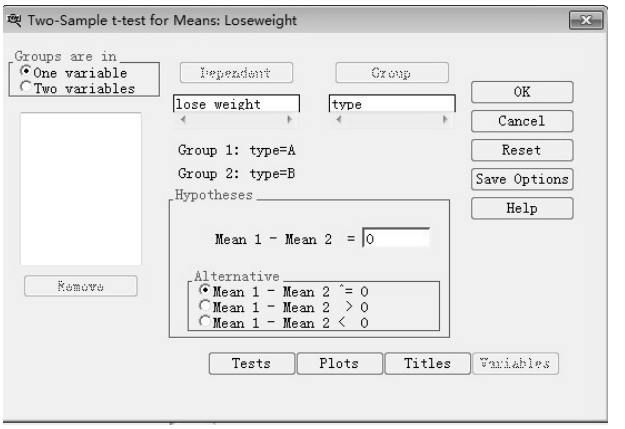


图 5-6　独立样本 T 检验对话框

若单击图 5-6 所示对话框中的 Tests（检验）按钮，在弹出的子对话框中可以设置计算两样本均值差的单侧和双侧置信区间，还可以单击 Power Analysis（功效分析）标签在此设置进行功效分析（Perform power analysis）；若单击 Plots（图形）按钮，在此可以设置输出 Box-&-whisker plot（箱形触须图）、Bar chart（盒装图）、Means plot（均值误差图）、t distribution plot（*t* 分布图）以直观地显示检验结果；若单击 Titles（标题）按钮，在弹出的子对话框中定义三类标题：Global（全局标题）；Two-Sample t（两样本 *t* 检验的标题）；Settings（设置标题：是否包括日期、页码、筛选信息）。

5.3.6　SAS 实例——检验某新药疗效是否显著

例 5-4　为了判断某种新药的疗效是否显著，选取 20 名患者进行药效试验。首先测定受试者血压值，在经过一周服药治疗后再测量他们的血压值，数据记录如表 5-10 所示（舒张压/mmHg，相应的 SAS 数据集在光盘中的存储路径为“data\chap5\bp”）。问该种降压药治疗是否有效？（α=0.05）

表 5-10　服用降压药前后的收缩血压记录　　单位：mmHg

患者序号	1	2	3	4	5	6	7	8	9	10
治疗前	114	117	155	114	119	102	140	91	135	114
治疗后	93	113	128	96	120	93	105	90	103	95

续表

患 者 序 号	11	12	13	14	15	16	17	18	19	20
治 疗 前	103	140	136	126	108	142	113	115	116	121
治 疗 后	99	132	130	121	103	132	114	104	110	112

解析：本实例为典型的配对实验设计第二种情形，即观察试验前后同一受试者同一指标的数值变化。因此采用配对样本 *t* 检验。根据题意，原假设为：服用降压药前后患者的收缩血压均值差异不显著，备择假设为：服用降压药前患者的收缩血压均值高于服用后。

编程法：

编写如下程序（其在光盘中的存储位置为“proc\chap5\bp”）：

```
Proc ttest data=chap5.bp sides=u;          /*调用 ttest 过程进行右侧检验*/
Paired before*after;                       /*定义配对变量 before 和 after*/
Run;
```

程序运行后，得到的结果如表 5-11 至表 5-13 所示。表 5-11 和表 5-12 分别为变量 difference=before-after 的描述性统计量和其均值、方差的 95%置信区间。表 5-13 为 *t* 检验结果，得到 *t* 检验统计量为 4.84，对应的 *P* 值小于 0.0001（小于显著性水平 0.05），则拒绝原假设，认为服用降压药前患者的血压明显高于服用后，即该降压药疗效显著。

表 5-11　描述性统计量

N	Mean	Std Dev	Std Err	Minimum	Maximum
20	11.4000	10.5352	2.3557	−1.0000	35.0000

表 5-12　均值和方差的 95%置信区间

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
11.4000	7.3266	Infty	10.5352	8.0119	15.3873

表 5-13　*t* 检验结果

DF	t Value	Pr > t
19	4.84	<0.0001

菜单法：

步骤一：选择菜单 Solutions|Analysis|Analyst 命令进入 Analyst 操作界面。

步骤二：选择 File|Open By SAS name 命令，单击打开数据集 chap5.bp。

步骤三：选择 Statistics|Hypothesis Test|Two Simple Paired t-test for mean 命令，弹出如图 5-7 所示对话框，单击变量 before，再单击 Group1（第一组）按钮，定义变量 before 为第一组，类似地将变量 after 确定为第二组（Group2）。在 Hypotheses（假设检验）选项框中可设定原假设和备择假设：在 Null（原假设）后面的选项框中填入 0，即设定配对样本均值差为 0；单击选择 Alternative（备择假设）下的选项 Mean（Group1−Group2）>0，即设置配对样本均值差大于零。单击 OK 按钮提交设置则将输出与编程方法一致的结果。

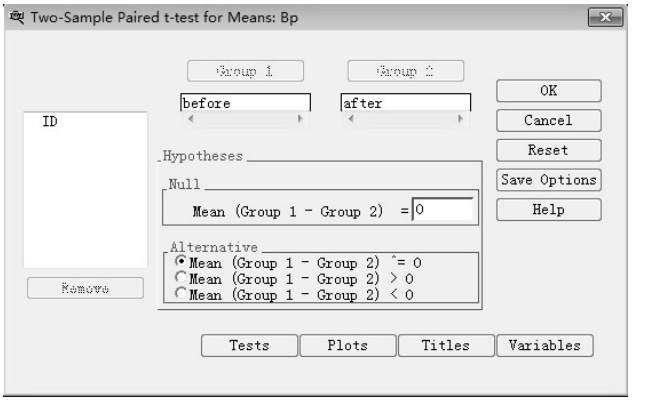


图 5-7　配对样本 *t* 检验对话框

5.3.7　SAS 实例——检验射击环数是否服从正态分布

例 5-5　已知某个射击队运动员一周内训练时的射中环数记录（共 735 条记录，部分数据如表 5-14 所示，包含完整数据的 SAS 数据集在光盘中的存储路径为“data\chap5\shooting”）。试分析该运动员的射击环数是否服从正态分布。

表 5-14　某运动员射击环数部分数据

7	7	8	8	8	9	9	7	8	9
6	8	7	8	7	9	8	7	7	7
8	7	10	9	8	8	7	7	8	6
9	8	7	9	8	8	10	7	6	9

解析：该实例为典型的正态分布检验问题，由于样本量为 735，小于 2000，则采用适用于小样本的 Shapiro-Wilk 的 *W* 检验。

编程法：

编写如下程序（其在光盘中的存储路径为“proc\chap4\shooting”）：

```
proc univariate data=chap5.shooting normal plot;      /*指定进行正态性检验并输出图形*/
var number;                                           /*指定对运动员射击环数（number）进行正态检验*/
run;
```

选择 Run|Submit 命令提交程序，正态分布检验结果如表 5-15 所示，观察可得第一行的 Shapiro-Wilk 检验对应的 *W* 统计量为 0.770398，对应的 *P* 值小于 0.0001，则拒绝“该运动员的射中环数服从正态分布”的原假设，同时由茎叶图、盒形图（如图 5-8 所示）和正态概率图（如图 5-9 所示）的分布综合判断得到一致结论。

菜单法：

步骤一：选择 Solutions|Analysis|Analyst 命令进入 Analyst 操作界面。

步骤二：选择 File|Open By SAS name 命令，单击选择打开数据集 chap5.shooting。

步骤三：选择 Statistics|Descriptive|Distributions 命令，弹出如 5-10 所示对话框，单击选中变量 number 为分析变量。

表 5-15　正态分布检验结果

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.770398	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.237702	Pr > D	<0.0100
Cramer-von Mises	W-Sq	8.167386	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	46.43184	Pr > A-Sq	<0.0050

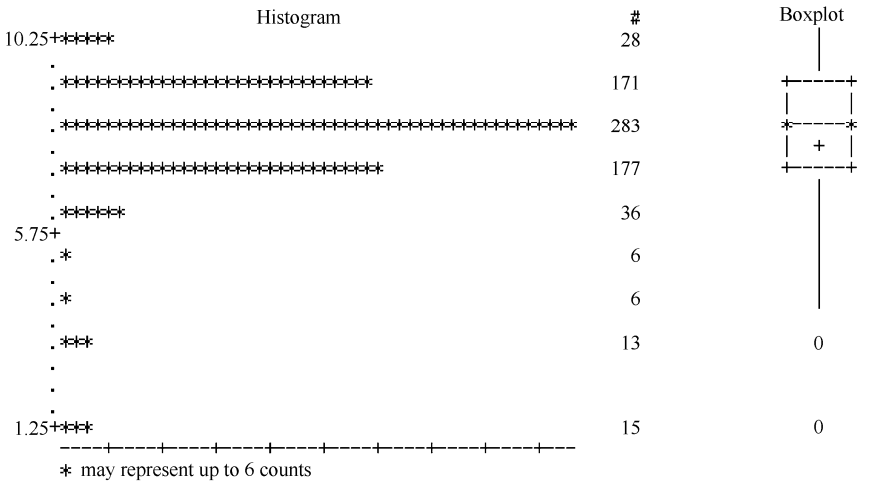


图 5-8　茎叶图和盒形图

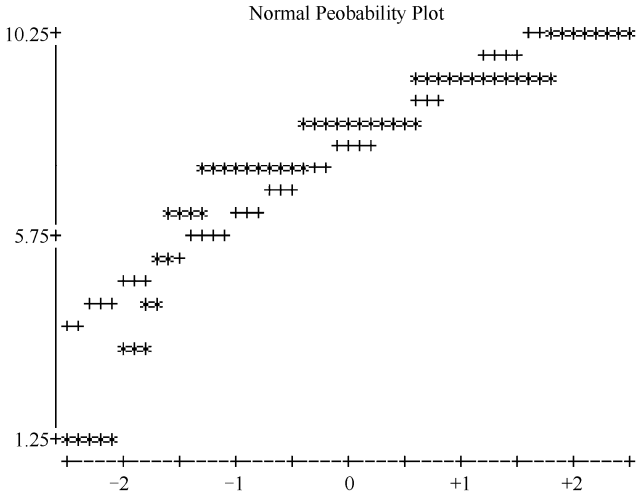


图 5-9　正态概率图

单击 Fit（拟合）按钮，弹出如图 5-11 所示对话框，Fit distributions（拟合分布）选项框内包含如下分布：Normal（正态分布）、Lognormal（对数正态分布）、Exponential（指数分布）、Weibull（威布尔分布），每个分布的参数（Parameters）可以采用系统默认设置（样本估计值）或自定义。在此选择正态分布，且采用样本估计值（Sample estimates）作为正态分布的参数

（Parameters）。单击 OK 按钮保存设置并返回如图 5-10 所示对话框。单击 OK 按钮将输出与编程法一致的结果。

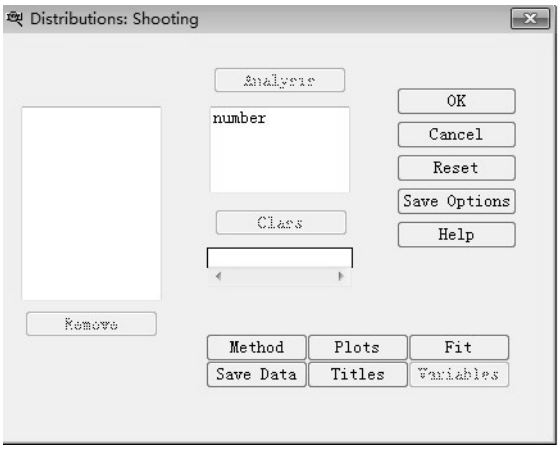


图 5-10　分布拟合检验对话框

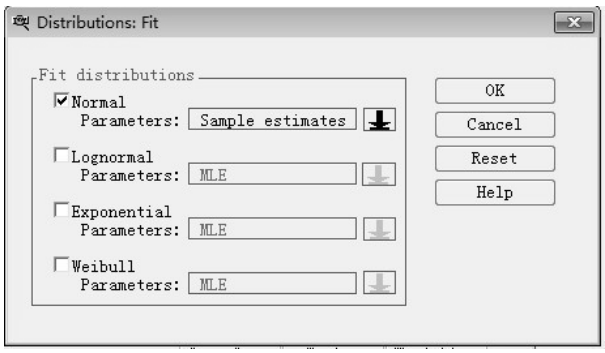


图 5-11　选择拟合类型

练习题

习题 5-1　某奶粉厂有一台盒装纯牛奶的机器，在正常生产时，每盒纯牛奶的净重服从正态分布，均值为 250g。为检查近期机器是否正常，从中抽取 20 盒牛奶，称得其净重数据如表 5-16 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap5\milk”）。请问在显著性水平为 0.01 时是否可以认为该机器生产正常？

表 5-16　盒装牛奶净重抽查数据　　单位：g

244	244	236	254	245	243	244	242	249	250
241	247	240	260	248	242	251	245	243	247
252	238	242	254	232	239	257	244	249	250

（本习题的解答程序在光盘中的存储路径为“proc\chap5\milk”。）

习题 5-2 为了了解食品的包装方式对其销售量是否存在影响，在某店将食品用两种不同的包装方式进行销售，得到一个月的销售情况（单位为件）如表 5-17 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap5\sales”）。请据此分析不同包装方式的销售量是否存在差异？

表 5-17 某食品不同包装方式的销售量 单位：件

Day（日）	方式一（Method1）	方式二（Method2）	Day（日）	方式一（Method1）	方式二（Method2）	Day（日）	方式一（Method1）	方式二（Method2）
1	113	79	11	129	110	21	144	125
2	154	97	12	128	81	22	152	110
3	96	90	13	159	113	23	127	96
4	91	79	14	90	66	24	133	129
5	117	144	15	122	88	25	129	64
6	108	77	16	148	112	26	101	149
7	117	95	17	199	128	27	90	76
8	136	156	18	148	120	28	146	127
9	160	148	19	112	74	29	110	108
10	177	88	20	148	86	30	121	95

（本习题的解答程序在光盘中的存储路径为“data\chap5\sales”。）

习题 5-3 某公司的营销部门在 20 家门店同时进行了为期一周的所有商品 8.8 折的促销活动，并且记录下了促销前一周和促销中一周的净利润（如表 5-18 所示，单位为元），相应的 SAS 数据集在光盘中的存储路径为“data\chap5\income”。请问促销后的利润相对于促销前是否增加了？（α=0.01）

表 5- 18 促销前后门店的净利润 单位：元

编号（ID）	促销前利润（Before）	促销后利润（After）	编号（ID）	促销前利润（Before）	促销后利润（After）
1	5952	6346	11	2360	2775
2	6787	7144	12	3505	3900
3	4323	4722	13	3075	3451
4	4948	5286	14	2358	2778
5	4493	4860	15	6567	6902
6	1683	2081	16	2187	2632
7	586	1004	17	3360	3711
8	6206	6614	18	7291	7708
9	2559	2907	19	6766	7158
10	3689	4053	20	2947	3322

（本习题的解答程序在光盘中的存储路径为“proc\chap5\income”。）

习题5-4 某人调查了从早上8:30到下午5:30班每隔1小时10个柜台的办理业务的顾客数，

如表 5-19 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap5\custom”）。请验证顾客数是否服从正态分布？

表 5- 19　每隔 1 小时 10 个柜台办理业务的顾客数

时间点	柜台 1	柜台 2	柜台 3	柜台 4	柜台 5	柜台 6	柜台 7	柜台 8	柜台 9	柜台 10
8:30～9:30	12	12	13	13	14	14	14	14	15	15
9:30～10:30	20	29	19	21	22	24	19	15	18	20
10:30～11:30	18	19	22	15	19	18	19	20	24	21
11:30～12:30	20	23	16	24	15	26	17	15	28	22
12:30～1:30	16	17	17	20	16	19	19	22	16	18
1:30～2:30	23	15	25	18	17	16	26	15	26	22
2:30～3:30	24	20	20	20	30	24	18	22	18	21
3:30～4:30	18	20	20	19	29	20	19	16	23	22
4:30～5:30	21	17	26	15	22	21	18	20	21	16

（本习题的解答程序在光盘中的存储路径为“proc\chap5\custom”。）

第6章 方 差 分 析

第 5 章简介了基本的参数估计和假设检验，着重介绍了单样本和两样本 *t* 检验，其中单样本 *t* 检验的目的是通过抽样的方式检验总体均值和某特定值的差异显著性，两样本 *t* 检验（包括配对样本 *t* 检验和独立样本 *t* 检验）的目的是比较两个正态总体均值的差异，若需要比较三个及三个以上正态总体均值的差异，则需要用到本章介绍的方差分析。

本章首先介绍在 SAS 系统中用于方差分析的 ANOVA 和 GLM 过程，然后结合常见实验设计方法介绍方差分析的原理，具体有单因素方差分析、区组设计方差分析、拉丁方设计方差分析、析因设计方差分析及协方差分析，最后结合 SAS 实例综合运用编程和菜单操作的方式完成方差分析。

6.1 SAS 过程——ANOVA 过程

在 SAS 系统中，均衡数据（分类变量每种组合中的观测数相等）的处理可应用 ANOVA 过程，非均衡数据的处理则应用 GLM 过程。ANOVA 过程的一般使用格式为：

```
PROC ANOVA      DATA=SAS 数据集 <选项列表> ;
CLASS           变量列表;
MODEL           因变量列表=自变量列表  </选项列表>;
MEANS           效应列表 </选项列表>;
RUN;
```

在 PROC ANOVA 语句后可使用的主要控制选项如表 6-1 所示。

表 6-1 PROC ANOVA 语句后主要控制选项

选　项	含　义
MANOVA	要求分析中删除因变量为缺失值的观测
NOPRINT	抑制所有的结果列表输出
OUTSTAT=	指定输出包括模型中每个效应的离差平方和、*F* 统计量和 *P* 值的输出数据集
PLOTS=NONE	抑制 ODS 图形输出

ANOVA 过程中的定义语句含义如下：

CLASS 语句——必须定义的分类变量，可为数值型或字符型变量，且必须放在 MODEL 语句前面。

MODEL 语句——用于定义因变量和自变量效应。若未指定自变量的效应，则只拟合截距。它的 4 种主要定义形式如表 6-2 所示。

表 6-2　MODEL 语句主要定义形式

形　　式	意　　义
MODEL Y=A B C;	主效应模型
MODEL Y=A B C　A*B　A*C　B*C　A*B*C;	含有交叉因素的模型
MODEL Y=A B C(A B);	嵌套模型
MODEL Y=A B(A) C(A) B*C(A);	包含嵌套、交叉和主效应的模型

在 MODEL 语句的斜杠（/）后选项列表中控制选项主要有：

INT/INTERCEPT——要求 ANOVA 过程把截距作为一个效应处理，输出与其有关的假设检验结果。ANOVA 过程在模型拟合时总是含有截距，但是若此选项默认，则不输出与之相应的假设检验结果。

NOINT——模型中不包含截距项。

NOUNI——不输出单变量分析结果。

MEANS 语句——计算此语句定义的每个效应所对应的因变量均值，MEANS 语句后选项列表中的主要控制选项有：选择多重比较的检验方法（如表 6-3 所示）及规定这些检验的细节（如表 6-4 所示），注意细节选项只能用于主效应。

表 6-3　多重比较方法列表

选　　项	含　　义
BON	对所有主效应均值之差进行 BONFERRONI 的 *t* 检验
DUNCAN	对所有主效应均值进行 DUNCAN 的多重极差检验
DUNNETT<(‘格式化对照值’)>	进行 DUNNETT 的双尾 *t* 检验，用以检验对所有主效应均值的某个水平作为对照，检验其他水平与之相比效应值差异的显著性。在括号内用单引号把对照效应的水平格式化值括起来。默认时，效应的第一个水平为对照
DUNNETTL<(‘格式化对照值’)>	DUNNETT 的单尾 *t* 检验，它检验是否任意一个处理显著地小于这个对照
DUNNETTU<(‘格式化对照值’)>	DUNNETT 的单尾 *t* 检验，它检验是否任意一个处理显著地大于这个对照
GABRIEL	对所有主效应均值进行 GABRIEL 多重对比检验
REGWF	对所有主效应均值进行 RYAN-EINOT-GABRIEL-WELSCH 多重 *F* 检验
REGWQ	对所有主效应均值进行 RYAN-EINOT-GABRIEL-WELSCH 多重极差检验
SCHEFFE	对所有主效应均值进行 SCHEFFE 的多重对比检验
SIDAK	对所有主效应均值水平依据 SIDAK 不等式进行调整后，对其均值之差两两进行 *t* 检验
SMM|GT2	当样本量不等时，基于学生化最大模和 SIDAK 不相关 T 不等式对主效应均值进行两两对比检验，当效应组样本量不等时，即等同为 HOCHBERG 的 GT2 方法
SNK	对所有主效应均值进行 STUDENT-NEWMAN-KEULS 多重极差检验
T|LSD	对所有主效应均值进行两两 *t* 检验，相当于在单元观察数相等时 FISHER 的最小显著差（FISHER’S LEAST-SIGNIFICANT-DIFFERENCE）检验
TUKEY	对所有主效应均值进行 TUKEY 的学生化极差检验
WALLER	对所有主效应均值进行 WALLER-DUNCAN 的 K 比率（K-RATIO）检验

表 6-4　多重比较检验细节可选项列表

选　项	含　义
ALPHA=P	给出均值间对比检验的显著性水平，默认值是 0.05
CLDIFF	把两两均值差的结果以置信区间的形式输出
CLM	对变量的每个水平的均值按置信区间形式输出
E=效应	指定在多重对比检验中所使用的误差均方。如果默认，使用残差均方（MS）。指定的效应必须是在 MODEL 语句中出现过的效应
KRATIO=值	给出 WALLER-DUNCAN 检验的类型 1/类型 2 的误差限制比例。KRATIO 的合理值为 50、100、500，大约相当于两水平时 ALPHA 值为 0.1、0.05、0.01。默认值为 100
LINES	降序列出所有检验方法产生的均值，并用一条线段在均值旁指出非显著的子集
HOVTEST	输出不同水平两两方差齐性的 LEVENE 检验

6.2　SAS 过程——GLM 过程

GLM（General Linear Models）即一般线性模型，它能应用于多种不同分析，如简单回归、多元回归、方差分析、协方差分析、加权回归、多项式回归、偏相关分析、多元方差分析等。在 GLM 过程中的大多数方差分析的语句和选项与 ANOVA 过程中基本相同，仅增加 CONTRAST、ESTIMATE 和 LSMEANS 语句。

GLM 过程的一般使用格式为：

```
PROC GLM    DATA=SAS 数据集名 <选项列表> ;
CLASS       变量列表;
MODEL       因变量列表=自变量列表 </选项列表>;
CONTRAST    ‘标签’ 效应 值表 </选项列表>;
LSMEANS     效应列表 </选项列表> ;
MEANS       效应列表 </选项列表> ;
OUTPUT      <OUT=输出数据集名> <统计量关键字=变量名列表>;
RUN;
```

使用过程中必须定义 CLASS 和 MODEL 语句，且 CLASS 语句必须出现在 MODEL 语句前，其他语句必须放在 MODEL 语句后。以下介绍 GLM 与 ANOVA 过程相比不同的和新增的语句。

MODEL 语句——GLM 过程的 MODEL 语句可使用几种不同效应，相应的例子如表 6-5 所示，表中 A、B、C 代表分类变量，Y_1、Y_2、X_1、X_2 代表连续变量。

表 6-5　不同形式 MODEL 语句的意义

形　式	意　义
MODEL Y=A;	单因素方差分析
MODEL Y=A B C;	主效应模型
MODEL Y=A B A*B;	因素模型
MODEL Y=A B(A) C(B A);	嵌套模型
MODEL Y1 Y2=A B;	多元方差分析模型
MODEL Y=A X1	协方差分析模型

MODEL 语句后主要的控制选项如表 6-6 所示（注意：与 ANOVA 过程的 MODEL 语句后使用的相同选项将不再列出）。

表 6-6　MODEL 语句主要选项

选　项	含　义
SOLUTION	输出参数估计值
E1/E2/E3/E4	输出模型中每一效应的类型 1/类型 2/类型 3/类型 4 的估计函数，并计算相应的平方和
SS1/SS2/SS3/SS4	对每个效应，输出与类型 1/类型 2/类型 3/类型 4 的估计函数相关的平方和
ALPHA=0.01/0.05/0.1	指定置信区间的α水平。系统默认值为 0.05
CLI/CLM	输出每一观察的预测值/预测均值的置信限，两者不能同时使用
P	输出自变量不含缺失值的每一个观测值、预测值和残差值，并输出 Durbin-Waston 统计量
XPX	打印叉积矩阵 $\boldsymbol{X'X}$
I	打印矩阵 $\boldsymbol{X'X}$ 的逆矩阵或广义逆矩阵

CONTRAST 语句——实现某些假设检验，如控制某些因素在某一水平时，对其他因素水平间做两两比较。CONTRAST 语句的写法较复杂，在使用过程中需特别注意，以下举一个简单例子：假设因素 A 有三个水平，因素 B 有两个水平，且这两个因素的交互作用显著，即 A 和 B 的不同水平组合所形成的不同试验条件对试验结果的影响较大。实验者关心两个因素分别取什么水平时，效果最好。以下控制 A 因素分别在 A1、A2、A3 水平上对 B 因素的三个水平进行两两均值比较，然后再控制 B 因素在 B1、B2 水平上，对 A 因素的两个水平间进行两两均值比较。

控制因素 A 的语句写法：

```
PROC GLM;
CLASS A B;
MODEL X= A B A*B;
CONTRAST 'B1 VS B2/A1' B 1 -1 0 A*B 1 -1;
CONTRAST 'B1 VS B3/A2' B 1 0 -1 A*B 0 0 1 -1;
CONTRAST 'B2 VS B3/A3' B 0 1 -1 A*B 0 0 0 0 1 -1;
RUN;
```

控制因素 B 的语句写法

```
PROC GLM;
CLASS B A;
MODEL X= B A B*A;
CONTRAST 'A1 VS A2/A1' A 1 -1 A*B 1 -1 0;
CONTRAST 'A1 VS A3/A2' A 1 -1 A*B 0 0 0 1 0 -1;
CONTRAST 'A2 VS A3/A3' A 1 -1 A*B 0 0 0 0 0 0 0 1 -1;
RUN;
```

LSMEANS 语句——计算列在本语句中的每一个效应的最小二乘均值（LSM）。最小二乘均值估计是针对非均衡数据设计的。LSMEANS 语句后主要的可选控制选项如表 6-7 所示。

表 6-7　LSMEANS 语句后的可选项

选　项	含　义
COV	在选项 OUT=指明的输出数据集中输出协方差
E	输出用以计算最小二乘均值的估计函数
E=效应	规定模型中的某个效应作为误差项
OUT=输出数据集名	产生一个包含 LSM 值、标准差及协方差的输出数据集
PDIFF	输出假设检验 $H_0: LSM(i) = LSM(j)$ 所有可能 P 值
STDERR	打印 LSM 的标准差和 $H_0: LSM = 0$ 的检验 P 值
TDIFF	打印假设检验 $H_0: LSM(i) = LSM(j)$ 的 T 值和相应的 P 值
ADJUST=BON/DUNNETT/SCHEFFE /SIDAK/SMM/GT2/TUKEY/T	要求对最小二乘均值之差的检验 P 值和置信限进行多重调整。默认值为 T
SLICE=效应	通过规定的这个效应来分开交叉的 LSM 效应。例如，假定交叉项 A*B 是显著的，如果想对 B 的每个效应检验 A 的效应，使用以下语句：LSMEANS　A*B　/SLICE=B;

6.3　单因素方差分析

6.3.1　基本原理

方差分析的目的主要是考察某一个因素（或者某几个因素的交互作用）是否对我们关注的变量产生显著的影响。它可被用于分析既定的数据，如考察 5 个年龄阶段（从 15～65 岁，每 10 年划分为一个阶段）人群的舒张血压是否有显著的差异；但多数情况下方差分析被用于先进行严谨的试验设计，通过试验获取的数据。因此首先简介试验设计的基本要素：

研究变量——又称因变量，是试验观察的主要指标。一次试验时可以记录下多个观察指标，相应的方差分析可同时设置多个因变量。

因素和水平——试验的因素可以是品种、人员、方法、时间、地区等，因素所处的状态称为水平。在每一个因素下面可以有若干水平。例如，某工厂的原料来自 4 个不同地区，检验用不同地区的原料生产的产品质量是否有显著差异，“地区”就是因素，4 个地区便是“地区”这一因素的 4 个水平。当某个主要因素的各个水平间的主要因变量的均值差异有显著的统计学意义时，必要时可进行两两水平间的比较。

因素间的交互作用——多因素的试验设计有时需要分析因素间的交互影响。两个因素间的交互影响称为一级交互影响，如因素 A 与因素 B 的一级交互影响可记为 A×B，三个因素间的交互影响称为二级交互影响，如因素 A 与因素 B 与因素 C 的二级交互影响可记为 A×B×C。当交互影响项呈现统计不显著时，表明各个因素独立；当呈现统计显著时，就需要列出这个交互影响项的效应，以助于做出正确的统计推断。

单因素方差分析按照试验设计时受试对象的抽取或分组的随机程度可分成两类：

完全随机设计——从符合条件的总体中完全随机地抽取所需数目的受试对象，再将全部受试对象完全随机地分配到 k 组中去。此时受试对象与试验因素间无直接联系。

组内完全随机设计——按照试验因素的 k 个水平将全部受试对象划分成 k 个子总体，再分别从 k 个子总体中完全随机地抽取所需数目的受试对象。此时试验因素的各个水平决定了受试对象各自应该归属的组别。

设因素 A 有 k 个水平 $A_1, A_2, \cdots, A_k$，在每一个水平下考察的指标可以看成一个总体，即 k 个水平代表了 k 个总体，现假定：

- 每一总体均服从正态分布；
- 每一总体的方差相同；
- 从每一总体中抽取的样本相互独立。

以下检验各总体的均值是否相同，设第 i 个总体的均值为 μ_i，则：

原假设 $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$，备择假设 $H_1: \mu_1, \mu_2, \cdots, \mu_k$ 不全相同

设从第 i 个总体获得样本量为 n_i 的样本观察值为 $y_{i1}, y_{i2}, \cdots, y_{in_i}$，$i = 1, 2, \cdots, k$，各样本间相互独立。样本观察值 y_{ij} 可看成是来自均值为 μ_i 的总体，这样 y_{ij} 就是其均值 μ_i 与随机误差 ε_{ij} 叠加加而产生的。上面我们已经假定在 A_i 水平下的 y_{ij} 服从 $N(\mu_i, \sigma^2)$ 分布，则有 $\varepsilon_{ij} \sim N(0, \sigma^2)$。因此单因素方差分析的统计模型如下：

$$\begin{cases} y_{ij} = \mu_i + \varepsilon_{ij} (i = 1, 2, \cdots, k \ , \ \ j = 1, 2, \cdots, n_i) \\ \text{各} \varepsilon_{ij} \text{相互独立，且都服从} N(0, \sigma^2)) \end{cases}$$

定义各个 μ_i 的加权平均 $\mu = \dfrac{1}{n} \sum_{i=1}^{k} n_i \mu_i$ 为总平均，其中 $n = \sum_{i=1}^{k} n_i$，于是 $a_i = \mu_i - \mu$（$i = 1, \cdots, k$）为因素 A 在第 i 水平的主效应，也简称为 A_i 的效应，同时也表明第 i 个总体的均值是一般平均与其效应的叠加。效应间有关系 $\sum_{i=1}^{k} n_i a_i = 0$。此时，单因素方差分析的统计模型可改写成：

$$\begin{cases} y_{ij} = \mu_i + a_i + \varepsilon_{ij} (i = 1, 2, \cdots, k \ , \ j = 1, 2, \cdots, n_i) \\ \sum_{i=1}^{k} n_i a_i = 0 \\ \text{各} \varepsilon_{ij} \text{相互独立，且都服从} N(0, \sigma^2) \end{cases}$$

所要检验的原假设也可改写成：

$$H_0: a_1 = a_2 = \cdots = a_k = 0$$

造成各 y_{ij} 间差异的原因可能有两个：一个可能是原假设 H_0 不真，即各水平下总体均值 μ_i（或水平效应 a_i）不同，因此从各总体中获得的样本观察值也就有差异了；另一个可能是 H_0 为真，差异是由于随机误差引起的。为了进一步定量分析这些差异，我们需要把这些差异表达出来。将组内观察值的平均值表述为：$\overline{y}_{i\cdot} = \mu_i + \overline{\varepsilon}_{i\cdot}$，其中 $\overline{y}_{i\cdot} = \sum_{j=1}^{n_i} y_{ij} / n_i$，$\overline{\varepsilon}_{i\cdot} = \sum_{j=1}^{n_i} \varepsilon_{ij} / n_i$，即组内样本观察值的平均值等于组内总体均值加上组内随机误差的平均值。将所有样本观察值的平均值表述为：$\overline{y} = \mu + \overline{\varepsilon}$，其中 $\overline{y} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} / n$，$\overline{\varepsilon} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \varepsilon_{ij}$，即所有样本观察值的平均值等于总平均（各组均值的加权平均）加上所有随机误差的平均值。于是每一个观察值 y_{ij} 与总平均 $\overline{y}$ 的离差可以分解成两部分：

$$y_{ij}-\bar{y}=(y_{ij}-\bar{y}_{i\cdot})+(\bar{y}_{i\cdot}-\bar{y})$$

其中 $y_{ij}-\bar{y}_{i\cdot}$ 称为组内离差，即得到：

$$y_{ij}-\bar{y}_{i\cdot}=(\mu_i+\varepsilon_{ij})-(\mu_i+\bar{\varepsilon}_{i\cdot})=\varepsilon_{ij}-\bar{\varepsilon}_{i\cdot}$$

说明组内离差仅仅反映了随机误差。而 $\bar{y}_{i\cdot}-\bar{y}$ 称为组间离差，代入得到：

$$\bar{y}_{i\cdot}-\bar{y}=(\mu_i+\bar{\varepsilon}_{i\cdot})-(\mu+\bar{\varepsilon})=a_i+\bar{\varepsilon}_{i\cdot}-\bar{\varepsilon}$$

说明第 i 组间离差除了反映随机误差外还反映了第 i 个水平的效应 a_i。

各 y_{ij} 间总的差异大小用总离差平方和 S_{T} 表示：

$$S_{\mathrm{T}}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2$$

由随机误差引起的数据间的差异用组内离差平方和表示，也称误差离差平方和 S_{e}：

$$S_{\mathrm{e}}=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2$$

由于组间离差除了随机误差外，还反映了效应的差异，故由于效应不同引起的数据差异可以用组间离差平方和表示，也称因素 A 的离差平方和 S_{A}：

$$S_{\mathrm{A}}=\sum_{i=1}^{k}n_i(\bar{y}_{i\cdot}-\bar{y})^2$$

将总离差平方和进行分解：

$$\begin{aligned}S_{\mathrm{T}}&=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot}+\bar{y}_{i\cdot}-\bar{y})^2\\&=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{y}_{i\cdot}-\bar{y})^2+2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})(\bar{y}_{i\cdot}-\bar{y})\\&=\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2+\sum_{i=1}^{k}n_i(\bar{y}_{i\cdot}-\bar{y})^2\\&=S_{\mathrm{e}}+S_{\mathrm{A}}\end{aligned}$$

其中 $\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})=0$。证明了总的差异=组内差异+组间差异。由于

$$\frac{1}{\sigma^2}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2=\frac{1}{\sigma^2}\sum_{j=1}^{n_i}(\varepsilon_{ij}-\bar{\varepsilon}_{i\cdot})^2\sim\chi^2(n_i-1)$$

又由 χ^2 分布的可加性可知：

$$\frac{S_{\mathrm{e}}}{\sigma^2}=\sum_{i=1}^{k}\left[\frac{1}{\sigma^2}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\cdot})^2\right]=\sim\chi^2\left(\sum_{i=1}^{k}(n_i-1)\right)=\chi^2(n-k)$$

还可证明，在 H_0 为真时，即各组效应 a_i 都为 0 时

$$\frac{S_{\mathrm{A}}}{\sigma^2}\sim\chi^2(k-1)$$

因此可采用统计量：

$$F=\frac{S_{\mathrm{A}}/(k-1)}{S_{\mathrm{e}}/(n-k)}\sim F(k-1,n-k)$$

来检验假设，当 $\frac{S_A/(k-1)}{S_e/(n-k)} < F_\alpha(k-1,n-k)$ 时，接受原假设；否则，拒绝原假设。

当经过 F 检验拒绝原假设，则表明因素 A 是显著的，即 k 个水平对应的指标均值不全相等，但是在因素的不同水平之间不一定存在显著差异。实际上，当方差分析的结论是因素 A 显著时，还需要我们进一步去确认因素 A 的哪些水平间的效应值有显著差异。同时比较任意两个水平均值间有无显著性差异称为多重检验，即要以显著性水平 α 同时检验以下 C_k^2 个假设：

$$H_0^{ij}:\mu_i=\mu_j \qquad i<j,\quad i,j=1,2,\cdots,k$$

均值间的多重比较根据所控制误差的类型和大小有许多具体方法，如成组比较 t 检验法、Bonforroni t 检验法等，以下具体介绍多重比较检验方法。

t 检验和 Bonforroni 检验——当考察因素有 k 个水平时，对任意两个水平均值间的差异的显著性检验用如下 t 统计量：

$$t_{ij}=\frac{\bar{y}_{i\cdot}-\bar{y}_{j\cdot}}{\sqrt{\frac{S_e}{n-k}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}}\sim t(n-k)$$

两两比较的次数共有 $m=C_k^2=k(k-1)/2$，因此，共有 m 个置信水平，每次比较的显著水平：t 检验的方法取 α，完成所有比较后的整体显著水平等于 $1-(1-\alpha)^m$，当比较次数 m 越大，试验误差就越大；而 Bonforroni 检验的方法取 α/m，完成所有比较后的整体显著水平等于 $1-(1-\alpha/m)^m<\alpha$，即最大试验误差率小于 α。

LSD 检验——既可以通过两两比较的显著水平的特定限制来控制最终的试验误差率，也可以通过两两比较的绝对差异界限来判别显著性。最容易想到的这个界限就是在两两比较中采用的 t 检验法而得到 Fisher 最小显著差（LSD）为：

$$\mathrm{LSD}_{ij}=t_{\frac{\alpha}{2}}(n-k)\sqrt{\frac{S_e}{n-k}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

当 $\left|\bar{y}_{i\cdot}-\bar{y}_{j\cdot}\right|\geqslant \mathrm{LSD}_{ij}$ 时，则 $P\leqslant\alpha$。

SNK 检验和 Duncan 检验——属于多级检验法，使用多级检验可以获得同时检验的更高效率。多级检验分为步长增加法和步长减少法，SAS 系统采用步长减少法。当因素有 k 个水平时，即有 k 个均值需要比较，检验步骤为：

- 将均值由大到小排队，即 $\bar{y}_{1\cdot}\geqslant\bar{y}_{2\cdot}\geqslant,\cdots,\geqslant\bar{y}_{k\cdot}$。
- 比较 $\bar{y}_{1\cdot}$ 与 $\bar{y}_{k\cdot}$ 是否有显著差异，此时跨度 $a=k$。若两者之间无显著差异，说明其他均值之差比它小的任何两个水平均值之间的差别也不显著，因此比较停止；反之继续进行下一步。
- 比较 $\bar{y}_{1\cdot}$ 与 $\bar{y}_{k-1\cdot}$、$\bar{y}_{2\cdot}$ 与 $\bar{y}_{k\cdot}$ 是否有显著差异。此时这两个比较的跨度 $a=k-1$。若两水平效应值差异不显著，则停止比较。若每一步都有不满足停止比较的对比组存在，最后应到达跨度为 2，以至所有需要比较的相邻两水平均值全部比较完。

多级检验在作每一级比较时，通过控制比较误差率 γ_a 的显著水平来实现其最终要控制的试验误差率。γ_a 在每一级比较时可能是不同的，它是跨度 a 和整体试验误差率 α 的函数，即 $\gamma_a=f(a,\alpha)$。另外要注意的是，γ_a 其实就是每一级比较时特定统计量分布的显著水平。常用的

两种方法是 SNK 检验和 Duncan 检验。它们的检验统计量为 q：

$$q_{ij}=\frac{\overline{y}_{i\cdot}-\overline{y}_{j\cdot}}{\sqrt{\frac{S_e}{2(n-k)}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}}\sim q(a,n-k)$$

其中，a 是 $\overline{y}_{i\cdot}$ 和 $\overline{y}_{j\cdot}$ 之间的跨度值，q 分布的自由度是 a 和 $n-k$，显著水平为 γ_a。SNK 检验和 Duncan 检验的区别主要在于 γ_a 取值：SNK 检验 $\gamma_a=\alpha$（注意当比较次数很大时，最大试验误差率将趋向于 1）；Duncan 检验 $\gamma_a=1-(1-\alpha)^{a-1}$。

多重比较方法的选择首先要注意每种方法适用的试验设计条件，其次要关心所要控制的误差类型和大小。例如，某因素有 10 个水平，若采用常规的 t 检验进行多重比较，共需要比较的次数为 $C_{10}^2=45$ 次，即使每次比较时都把第一类错误 α 控制在 0.05 水平上，但经过 45 次多重比较后，犯第一类错误的概率上升到 $1-(1-0.05)^{45}=0.90$。从中我们可以看到选用 t 检验法进行多重比较，仅仅控制了每次比较的显著水平，却大大增加了整体的显著水平。

下面是所要控制的几种误差类型和选用的检验方法：

- 第一类误差率——即犯第一类错误的概率 α。
- 比较误差率——即每一次单独比较时，所犯第一类错误的概率，可使用 T 法、LSD 法、DUNCAN 法。
- 试验误差率——即完成全部比较后，整体所犯第一类错误的概率。
- 完全无效假设下的试验误差率——即在原假设完全无效下的试验误差率，可使用 SNK 法。
- 部分无效假设下的试验误差率——即在原假设部分无效下的试验误差率。
- 最大试验误差率——即在原假设完全或部分无效下，完成全部比较后所犯第一类错误的最大概率，可使用 BON 法、SIDAK 法、SCHEFFE 法、TUKEY 法、GT2/SMM 法、GABRIEL 法、REGWQ 法、REGWF 法、DUNNETT 法。

6.3.2 SAS 实例——2009 年不同地区商品房销售差异分析

例 6-1 已知 2009 年我国中部、东部和西部不同省市的商品房销售面积（万平方米）和商品房销售额（亿元）数据，如表 6-8 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap4\area”）。试分析不同地区的商品房销售面积是否存在显著的差异？

表 6-8 2009 年我国不同地区商品房销售情况

地　区	商品房销售面积（万平方米）	商品房销售额（亿元）
一、东部地区		
北　京	2362.25	3259.66
天　津	1590.02	1094.85
河　北	2849.14	941.83
辽　宁	5375.07	2168.29
上　海	3372.45	4330.22
江　苏	9922.73	4955.42
浙　江	5525.38	4302.98

续表

地　区	商品房销售面积（万平方米）	商品房销售额（亿元）
一、东部地区		
福　建	2723.23	1478.22
山　东	6931.7	2436.46
广　东	7035.89	4585.93
海　南	560.34	351
二、中部地区		
山　西	1014.39	275.77
吉　林	1823.22	540.26
黑龙江	2015.53	652.52
安　徽	4053.92	1378.39
江　西	2280.91	602.8
河　南	4338.6	1156.6
湖　北	2718.3	959.88
湖　南	3513.72	941.6
三、西部地区		
内蒙古	2463.01	733.22
广　西	2383.76	777.17
重　庆	4002.89	1377.76
四　川	5888.67	2074.91
贵　州	1619.25	467.81
云　南	2229.95	653.53
西　藏	14.23	4.73
陕　西	2086.97	672.72
甘　肃	696.26	174.59
青　海	218.32	54.9
宁　夏	775.29	239.53
新　疆	1327.64	351.01

解析：本实例为典型的单因素方差分析情形，其中因素为“地区”，因素取三个水平，分别为“中部地区”、“东部地区”和“西部地区”。注意到这三个地区包括的城市数目不一样，即不满足平衡设计的条件，因此采用 GLM 过程分析。

编程法：

编写如下程序（其在光盘中的存储路径为“proc\chap6\area”）：

```
proc glm data=chap6.area;          /*调用 glm 过程*/
class area;                        /*定义分类变量*/
model square_meter=area;           /*定义模型因变量为 square_meter，自变量为 area*/
means area/ hovtest Duncan;        /*计算不同区域的商品房出售面积均值；进行组间的方差齐性检验；
                                    应用 Duncan 多重比较方法*/
run;
```

选择 Run|Submit 命令提交程序。以下分析主要输出结果。

表 6-9 为模型方差分析表，模型显著性 *F* 检验对应的 *P* 值为 0.0288（<0.05），可知模型显著成立。表 6-10 所示为模型拟合统计量：判定系数（R-Square=0.223918）、变异系数（Coeff Var=68.20453）、均方根（Root MSE=2061.824）、因变量 square_meter 的均值（square_meter Mean=3032.001）。表 6-11 和表 6-12 为效应的方差分析表，两者在于离差平方和的计算方差有所区别，但在本例中两者的计算结果没有差别：效应 area 对应的 *P* 值为 0.0288，联系本实验背景可判断至少两个地区的商品房的销售面积有显著差异。

表 6-9　模型方差分析表

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	34343355.6	17171677.8	4.04	0.0288
Error	28	119031279.3	4251117.1		
Corrected Total	30	153374634.9			

表 6-10　模型拟合统计量

R-Square	Coeff Var	Root MSE	square_meter Mean
0.223918	68.20453	2061.824	3023.001

表 6-11　效应方差分析表（第一类平方和）

Source	DF	Type I SS	Mean Square	F Value	Pr > F
area	2	34343355.55	17171677.78	4.04	0.0288

表 6-12　效应方差分析表（第三类平方和）

Source	DF	Type III SS	Mean Square	F Value	Pr > F
area	2	34343355.55	17171677.78	4.04	0.0288

表 6-13 为列文方差齐性的 *F* 检验结果。检验 *P* 值（0.0741）大于显著性水平（0.05），则可接受原假设，认为各组样本的方差是相等的，即满足方差分析的前提条件。若不满足方差齐性的前提条件，可以考虑采用数据变换法。

表 6-13　Leven's 方差齐性检验结果

Levene's Test for Homogeneity of square_meter Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
area	2	2.007E14	1.004E14	2.86	0.0741
Error	28	9.828E14	3.51E13		

表 6-14 为 Duncan 多重检验结果：Duncan Grouping（Duncan 分组）标题下的不同字母（A、B、C）表示组别，Mean（均值）列为各组的均值降序排列，N 列为各组的样本量，area（标题）列为因变量 area 对应的水平。由此表可得出结论：东部地区和中部地区可以为一组，中部地区

和西部地区可作为另一组。

表 6-14　Duncan 多重检验结果

Means with the same letter are not significantly different.				
Duncan Grouping		Mean	N	area
	A	4386.2	11	东部地区
	A			
B	A	2719.8	8	中部地区
B				
B		1975.5	12	西部地区

联系实例背景可得：由列文方差齐性检验结果和模型的显著性检验可知用方差分析模型分析此问题是合理的；对效应变量 area 的显著性检验结果得到至少有两个地区的商品房销售面积有显著差异；Duncan 多重检验结果显示东部地区的商品房销售面积最大，中部地区次之，西部最小。东部地区和中部地区的商品房销售面积没有显著的差异，中部地区和西部地区的商品房销售面积也没有显著差异，但是东部地区和西部地区的商品房销售面积差异显著。

菜单法：

步骤一：选择 Solutions|Analysis| Analyst 命令，进入 Analyst 分析界面。

步骤二：选择 File|Open|Open By Sas name|chap6|area|OK 命令，打开数据集 chap6.area。

步骤三：选择 Statistics|ANOVA|One-Way ANOV 命令，弹出如图 6-1 所示对话框，单击定义因变量（Dependent）为 square_meter，自变量（Independent）为 area。

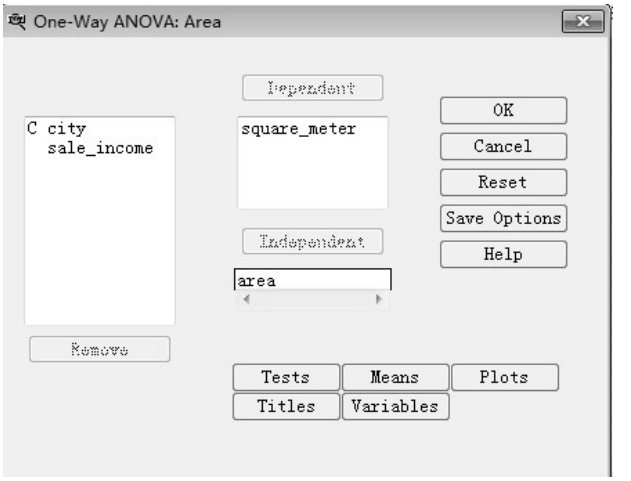


图 6-1　单因素方差分析主对话框

步骤四：单击 Tests（检验）按钮，弹出如图 6-2 所示对话框，选择 Levene's test 选项，即定义列文方差齐性检验。单击 OK 按钮保存设置并返回如图 6-1 所示对话框。

若单击 Power Analysis（功效检验）标签，则在相应的选项卡中可以设置相应的显著性水平进行检验功效分析。

步骤五：单击 Means（均值）按钮，弹出如图 6-3 所示对话框，单击 Comparison method（比较方法）选项框右侧的按钮，单击下拉菜单中的 Duncan's multiple-range test（邓肯多重

比较）选项；Significance level（显著性水平）采用系统默认的 0.05；单击变量 area，再单击 Add（加入）按钮，将 area 选进 Effect/method（效应）选项框。以上操作设置在 0.05 的显著性水平之下用邓肯多重比较法两两比较地区的商品房销售面积的差异。单击 OK 按钮保存设置并返回如图 6-1 所示对话框。

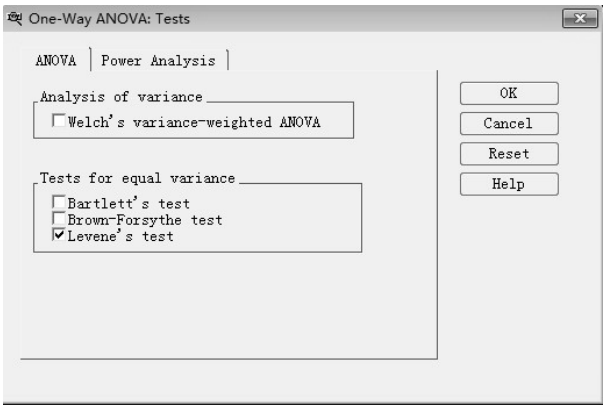


图 6-2　方差齐性检验

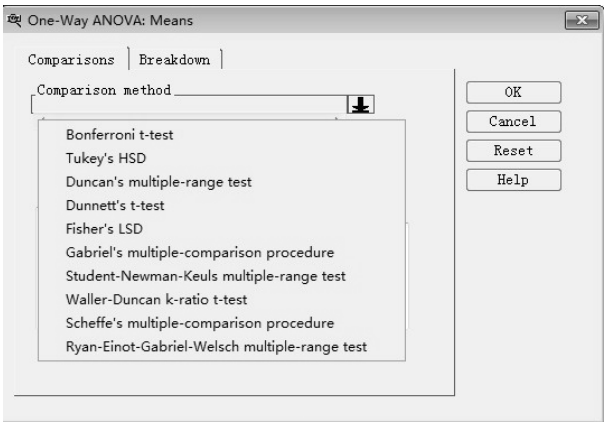


图 6- 3　多重比较

单击图 6-1 所示对话框上的 OK 按钮，除输出与方法一同样的结果外，还增加了对每个类型计算的因变量的均值和标准差。

作为练习，读者可以尝试用编程和菜单操作的方式比较不同地区的商品房销售价格的差异。

6.4　区组设计方差分析

6.4.1　基本原理

随机区组设计的步骤为：先将受试对象按性质相同或相近组成单位组，每个单位组包含 k 个受试对象；再将这些受试对象随机分配到因素 A 的 k 个水平上。由于每个水平受试对象数量

相同、性质相同或相近，以此减小误差，提高试验效率。若将单位组也看作一个因素，则此为两因素（因素 A 和单位组）设计，两个因素的各个水平仅交叉 1 次，此时一般不考虑交互效应。随机单位组设计的方差分析表的一般形式如表 6-15 所示。

表 6- 15　区组设计方差分析表

变 异 来 源	离差平方和	自　由　度	均　　方	F 统 计 量	*P*　　值
因素 A	SS_A	$k-1$	$MS_A= SS_A/(k-1)$	$F_A=MS_A/MS_e$	P_A
单位组	$SS_{单}$	$b-1$	$MS_{单}= SS_{单}/(b-1)$	$F_{单}=MS_{单}/MS_e$	$P_{单}$
误差 S_e	SS_e	$bk-k-b+1$	$MS_e= SS_e/(bk-k-b+1)$		
总变异 S_T	SS_T	$bk-1$	$MS_T= SS_T/(bk-1)$	$F_T=MS_T/MS_e$	P_T

6.4.2　SAS 实例——检测不同化学试剂对布匹强度影响的差异性

例 6-2　某化学家想要检验 4 种化学试剂在一种特定类型布料的强度上的效应。因为不同的布匹之间可能存在变异性，他采用随机化区组设计并将布匹考虑为区组。他选取了 4 匹布，并用所有 5 种试剂随机对每匹进行试验，测得的抗压强度数据如表 6-16 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap6\strength”）。分析数据并得到相应的结论。

表 6-16　布匹抗压强度数据

布匹（cloth）	化学试剂（reagent）				
	1	2	3	4	5
A	74	68	75	71	68
B	73	67	76	73	73
C	76	69	72	72	66
D	72	63	74	69	68

解析：本例的研究目的是比较不同的化学试剂对布匹的抗压强度的影响，因此进行方差分析。而且由于本实验为均衡设计，因此选择 ANOVA 过程。

编程法：

编写程序如下所示（其在光盘中的存储路径为“proc\chap6\strength”）：

```
proc anova data=chap6.strength;          /*调用 anova 过程*/
class reagent cloth;                     /*定义分组变量*/
model strength=reagent cloth;            /*定义模型*/
means reagent/tukey alpha=0.01;          /*用 tukey 方法两两比较不同化学试剂作用后的布匹强度，指
                                           定显著性水平为 0.01*/
means reagent;                           /*计算不同化学试剂作用后的布匹强度的均值和标准差*/
run;
```

选择 Run|Submit 命令提交程序，以下分析主要的输出结果。

由表 6-17 可知对模型的显著性 *F* 检验 *P* 值为 0.0020（< 0.05），则拒绝原假设，认为模型显著成立。表 6-18 列出了 4 个模型拟合参数，其中模型的拟合优度（R-Square）为 0.00795，说明

模型拟合较合适。观察效应的方差分析表（如表 6-19 所示）：自变量 reagent 对应的 *F* 检验 *P* 值为 0.0007（<0.05），自变量 cloth 对应的 *F* 检验 *P* 值为 0.1412（>0.05）。由此判断不同布匹的强度之间没有显著的差异，但是经过了不同的化学试剂处理后的布匹强度之间有显著的差异。

表 6-17　模型方差分析

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	191.3500000	27.3357143	6.89	0.0020
Error	12	47.6000000	3.9666667		
Corrected Total	19	238.9500000			

表 6-18　模型拟合参数

R-Square	Coeff Var	Root MSE	strength Mean
0.800795	2.807117	1.991649	70.95000

表 6-19　效应方差分析表

Source	DF	Anova SS	Mean Square	F Value	Pr > F
reagent	4	165.2000000	41.3000000	10.41	0.0007
cloth	3	26.1500000	8.7166667	2.20	0.1412

表 6-20 为 Tukey 多重比较的结果：5 种化学试剂（reagent 列）被分成了两组，第 1、第 3、第 4、第 5 种试剂被分到了 A 组，试剂 4、5 和 2 被分到 B 组。注意本表是按照布匹的强度均值降序排列的。联系本实验背景得出结论：使用第 3 种试剂作用后布匹的强度最佳，使用第 1、第 3、第 4 和第 5 种化学试剂作用后，布匹的强度没有显著差异，而使用第 2、第 4 和第 5 种化学试剂后布匹强度没有显著差异。即第 1、第 3 种试剂作用后布匹的强度显著地高于第 2 种试剂作用后。

表 6-20　Tukey 多重比较结果

Means with the same letter are not significantly different.				
Tukey Grouping		Mean	N	reagent
	A	74.250	4	3
	A			
	A	73.750	4	1
	A			
B	A	71.250	4	4
B	A			
B	A	68.750	4	5
B				
B		66.750	4	2

表 6-21 为不同的化学试剂作用以后对应的布匹强度的均值（Mean）和标准差（Std Dev）。

表 6-21　不同化学试剂作用对应布匹强度均值和标准差

Level of reagent	N	strength	
		Mean	Std Dev
1	4	73.7500000	1.70782513
2	4	66.7500000	2.62995564
3	4	74.2500000	1.70782513
4	4	71.2500000	1.70782513
5	4	68.7500000	2.98607881

菜单法：

步骤一：选择 Solutions|Analysis| Analyst 命令，进入 Analyst 分析界面。

步骤二：选择 File|Open|Open By Sas name|chap6|strength|OK 命令，打开数据集 chap6.strength。

步骤三：选择 Statistics|ANOVA|Linear Models 命令，弹出如图 6-4 所示对话框，单击选择变量 strength 为 Dependent（因变量），定义分类变量（class）为 reagent 和 cloth。

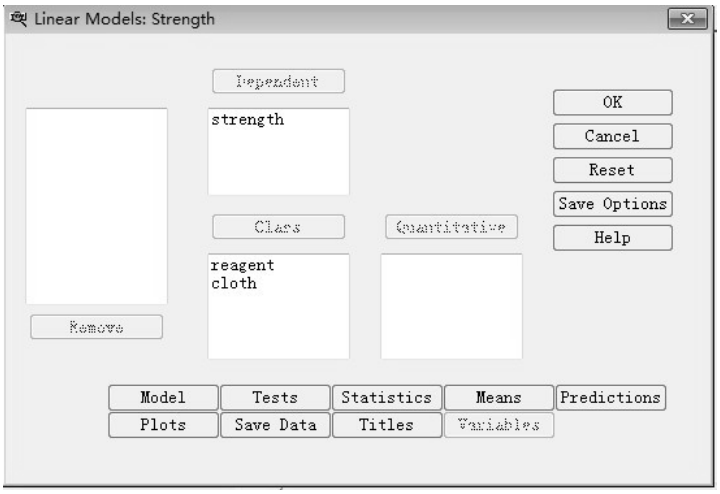


图 6-4　方差分析主对话框

步骤四：单击 Model（模型）按钮，弹出如图 6-5 所示对话框，单击 Standard Models（标准模型）按钮，单击选择下拉菜单中的 Main effects only（主效应），则变量 reagent 和 cloth 自动被选进 Effects in model（模型效应）选项框被定义为模型主效应。单击 OK 按钮保存设置并返回如图 6-4 所示对话框。

步骤五：单击 Means（均值）按钮，弹出如图 6-6 所示对话框，单击 Comparison method（比较方法）选项框旁边的按钮，在下拉菜单中单击选择 Tukey's HSD（Tukey's HSD 多重检验法）；Significance level（显著性水平）的值采用系统默认的 0.05；单击变量 reagent，再单击 Add（加入）按钮。以上操作设定采取 Tukey's HSD 方法对不同化学试剂（reagent）的布匹的强度（strength）均值进行两两比较。单击 OK 按钮保存设置并返回如图 6-4 所示对话框，单击 OK 按钮，将输出与编程方法同样的结果。

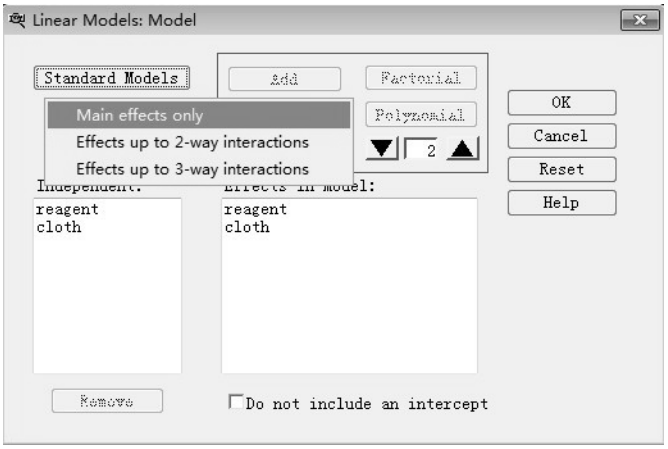


图 6-5　模型设定

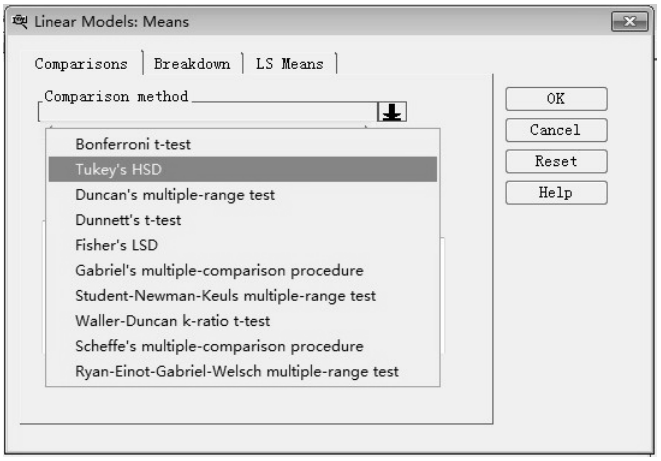


图 6-6　设定多重比较方法

6.5　拉丁方设计方差分析

6.5.1　基本原理

若试验中涉及三个因素，当它们之间不存在交互作用或交互作用可以忽略不计，即一般是包含一个试验因素和两个取相同水平数的区组因素，适合于拉丁方设计。用 K 个拉丁字母排成 K 行 K 列的方阵，使每行每列中每个字母仅出现一次，然后将区组因素放置到拉丁方的行、列上，将试验因素放置在字母上。例如，三个 4×4 的拉丁方为：

A	B	C	D		A	B	C	D		A	B	C	D
B	A	D	C		B	A	D	C		D	C	B	A
D	C	B	A		C	D	A	B		B	A	D	C

C　D　A　B　　D　C　B　A　　C　D　A　B

以下还列出了 4 个 5×5 的拉丁方为：

A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
B	C	D	E	A	C	D	E	A	B	D	E	A	B	C	E	A	B	C	D
C	D	E	A	B	E	A	B	C	D	E	A	B	C	D	D	E	A	B	C
D	E	A	B	C	B	C	D	E	A	B	C	D	E	A	C	D	E	A	B
E	A	B	C	D	D	E	A	B	C	C	D	E	A	B	B	C	D	E	A

在实际使用时，可以选择任何一个。拉丁方实验设计的方差分析应用离差平方和分解的思想，遵循方差分析一般原理。

6.5.2　SAS 实例——研究不同电视组装方法的组装时间的差异性

例 6–3　一位工业工程师研究 4 种组装方法（A、B、C、D）对彩色电视组建组装时间的效应。选取 4 位操作工来操作，采用如下所示的拉丁方设计，记录数据如表 6-22 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap6\assemble”）。试分析不同的组装方法所需要的组装时间是否有显著的差异？哪种组装方法最优？（$\alpha = 0.05$）

表 6-22　不同组装方法所需组装时间

组装顺序（order）	操作工（worker）			
	1	2	3	4
1	C=10	D=14	A=7	B=8
2	B=7	C=18	D=11	A=8
3	A=5	B=10	C=11	D=9
4	D=10	A=10	B=12	C=14

解析：由于本例的拉丁方设计采取了均衡实验设计，以下选用 ANOVA 过程进行方差分析。

编程法：

编写如下程序（其在光盘中的存储位置为“data\chap6\assemble”）：

```
proc anova data=chap6.assemble;                  /*调用 anova 过程*/
class worker methods order ;                     /*定义分类变量*/
model time=worker methods order;                 /*定义模型因变量为 time，自变量为 worker,
                                                   methods,order*/
means methods worker/snk alpha=0.05;             /*将因变量按 methods 分类进行均值比较，并设定
                                                   显著性水平为 0.05*/
run;
```

选择 Run|Submit 命令提交程序，以下分析主要输出结果。

由表 6-23 可知模型的 *F* 检验 *P* 值为 0.0072（<0.05），则方差模型显著成立。模型拟合的判别系数（R-Square）为 0.931373（如表 6-24 所示），则模型拟合度高。由表 6-25 所示效应方差分析表可知：因变量 time 按变量 methods 或 worker 分组的均值间差异显著，即采用不同的组装方法或由不同的操作者组装电视的时间存在显著差异。

表 6-23　模型的方差分析表

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	142.5000000	15.8333333	9.05	0.0072
Error	6	10.5000000	1.7500000		
Corrected Total	15	153.0000000			

表 6-24　模型拟合信息

R-Square	Coeff Var	Root MSE	time Mean
0.931373	12.90610	1.322876	10.25000

表 6-25　效应的方差分析表

Source	DF	Anova SS	Mean Square	F Value	Pr > F
worker	3	51.50000000	17.16666667	9.81	0.0099
methods	3	72.50000000	24.16666667	13.81	0.0042
order	3	18.50000000	6.16666667	3.52	0.0885

表 6-26 为用 SNK 多重比较方法两两比较不同组装方法（methods）对应的组装时间差异。将 4 种方法分成了三组：方法 C、D 为 A 组，方法 D、B 为 B 组、方法 B、A 为 C 组。在 0.05 的显著性水平下，组内的电视组装平均时间无显著差异，组间电视组装平均时间差异有统计学意义。

表 6-26　SNK 多重比较（methods）

Means with the same letter are not significantly different.				
SNK Grouping		Mean	N	methods
	A	13.2500	4	C
	A			
B	A	11.0000	4	D
B				
B	C	9.2500	4	B
	C			
	C	7.5000	4	A

表 6-27 为用 SNK 多重比较方法两两比较不同工作者（worker）对应的组装时间的差异。将 4 个操作者分成两组，2 号操作者为 A 组，1、3、4 号操作者为 B 组。则仅 2 号操作者操作时间明显（显著性水平为 0.05）长于其他操作者。

表 6-27　SNK 多重比较（worker）

Means with the same letter are not significantly different.			
SNK Grouping	Mean	N	worker
A	13.0000	4	2
B	10.2500	4	3
B			
B	9.7500	4	4
B			
B	8.0000	4	1

菜单法：

步骤一：选择 Solutions|Analysis| Analyst 命令，进入 Analyst 分析界面。

步骤二：选择 File|Open|Open By Sas name|chap6|Assemble|OK 命令，打开数据集 chap6.Assemble。

步骤三：选择 Statistics|ANOVA|Linear Model 命令，弹出如图 6-7 所示对话框，将变量 time 选定为因变量（Dependent）。类似的，选定分类变量（Class）为 worker、order 和 methods。

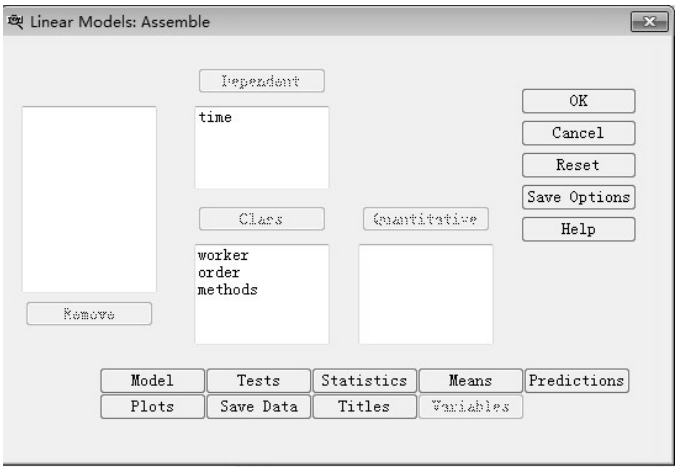


图 6-7　方差分析主界面

步骤四：单击 Model（模型）按钮，弹出如图 6-8 所示对话框，单击 Standard Models（标准模型）按钮，在下拉菜单中单击 Main effects only（主效应），则变量 worker、order 和 methods 被选进 Effects in medel（模型效应）选项框，即被定义为模型主效应。单击 OK 按钮保存设置并返回如图 6-7 所示对话框。

步骤五：单击 Means（均值）按钮，弹出如图 6-9 所示对话框，单击 Comparison method（比较方法）选项框旁的按钮，单击选择下拉菜单中的 Dunnett's t-test（Dunnett's *t* 检验法）；Significance level（显著性水平）采用默认的 0.05；单击变量 worker，再单击 Add（加入）按钮将其选进 Effect/method（效应/方法）选项框，类似的，将变量 methods 选进 Effect/method 选项框。以上操作设定采用 Dunnett's *t* 检验法两两比较用不同的方法（操作者）组装电视的时间的显著差异。单击 OK 按钮保存设置并返回主对话框（如图 6-7 所示）。单击 OK 按钮，则输出与

编程方法同样的结果。

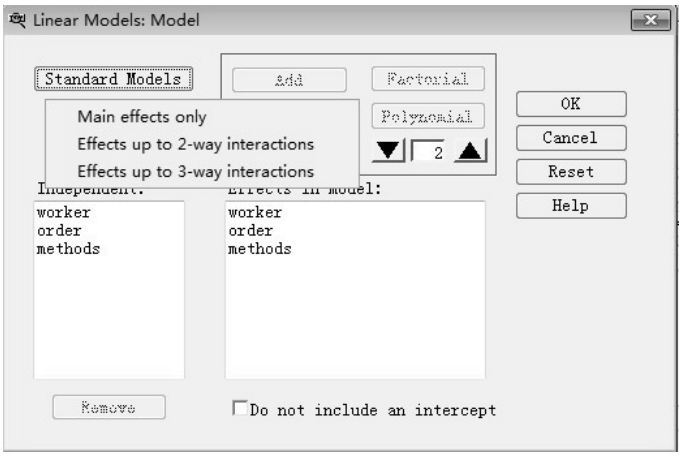


图 6-8　定义分析模型

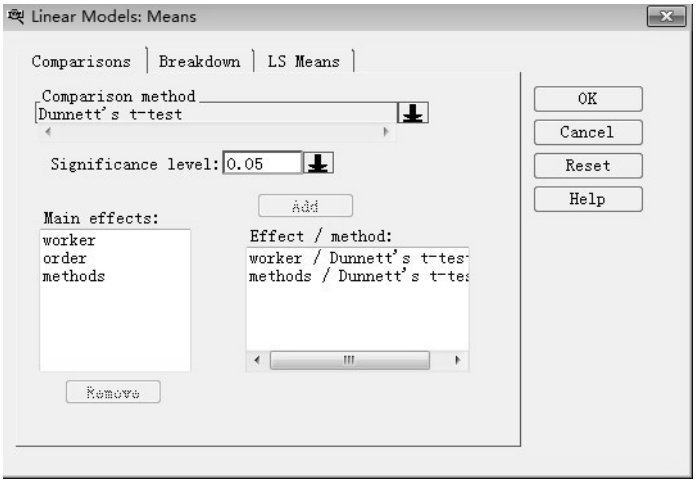


图 6-9　定义多重比较

6.6　析因设计方差分析

6.6.1　基本原理

析因设计是一种多因素实验设计方法，各因素在试验中所处的地位基本平等，而且因素之间存在一级（即两因素之间）、二级（即三因素之间）乃至更复杂的交互作用。例如，实验设计包括两个因素时，一个因素有三个水平，另一个因素有两个水平，全部水平组合共有 3×2=6 种组合，每种组合都做试验时就是析因试验设计。在每一种组合下，重复试验次数称为重复数。重复数可以不相等，但是一般而言重复数相等时效率最高。

析因设计不仅能够检验每个因素的各个水平间效应均值的统计差异，也能检验因素间的交互影响。因素间的交互作用指一个因素各个水平间的效应均值差异会随着另一个因素的水平改变而不同；反之，如果因素之间不存在交互影响时，一个因素的水平改变不影响另一个因素的各个水平的效应。

析因设计的方差分析的基本原理仍然为离差平方和的分解。例如，对 A×B 两个因素方差分析，总变异被分解成：A 因素的各个水平之间的差异、B 因素的各个水平之间的差异、A 与 B 的各种不同组合之间的差异及观察数据必然会产生的随机误差 4 部分。方差分析的主要目的就是要将这 4 部分从总平方和中分离出来，再以各个平方和与误差平方和进行比较。假设 A 因素有 x 个水平，B 因素有 y 个水平，每一种水平下的重复数为 k，那么总的观察数据有 $n=x\times y\times k$ 个，方差分析表如表 6-28 所示。

表 6-28　双因素（$r\times c$）重复数 m 的方差分析表形式

变异 source	离差平方和 SS	自由度 df	均方 MS	F 统计量 F	P 值 P
因素 A	SS_A	$x-1$	$MS_A=SS_A/(x-1)$	$F_A=MS_A/MS_e$	P_A
因素 B	SS_B	$y-1$	$MS_B=SS_B/(y-1)$	$F_B=MS_B/MS_e$	P_B
A×B	SS_{AB}	$(x-1)(y-1)$	$MS_{AB}=SS_{AB}/((x-1)(y-1))$	$F_{AB}=MS_{AB}/MS_e$	P_{AB}
误差 S_e	SS_e	$x\times y\times(k-1)$	$MS_e=SS_e/(xy(m-1))$		
总变异 S_T	$SS_T=SS_A+SS_B+SS_{AB}+SS_e$	$x\times y\times k-1$	$MS_T=SS_T/(xyk-1)$	$F_T=MS_T/MS_e$	P_T

6.6.2　SAS 实例——研究温度和压强对某化学物品产率的影响

例 6-4　在研究化学过程的产率过程中，两个最主要的变量是压强和温度。每一个因素选取三个水平，进行有 4 个重复的析因实验。产率数据如表 6-29 所示，相应的 SAS 数据集在光盘中的存储路径为“data\chap6\outcome”。试分析温度和压强对化学物品产率的影响，考察这两个因素的交互作用是否对化学物品产率有影响？

表 6- 29　某化学物品产率记录

温　　度 （Temp）	压强（pressure）					
	200Pa		215Pa		230Pa	
A：低（≤20℃）	90.4	90.5	94.8	93.6	89.2	88.1
	90.1	90.4	93.7	94.6	90.4	90.2
B：中（>20℃且≤30℃）	91.1	91.2	93.5	93.6	90.9	90.1
	91.3	90.9	93.4	93.3	90.7	90.3
C：高（>30℃）	92.5	92.6	92.7	92.5	91.4	91.3
	92.7	92.3	92.6	92.4	91.3	91.1

解析：本例将采用 GLM 过程进行分析，并尝试应用实现控制其中一个因素在某一水平上，对另一个因素的各个水平进行两两比较的 Contrast 语句。

编程法：

编写如下程序（其在光盘中的存储路径为“proc\chap6\outcome”）：

（1）首先编写如下程序粗略考察温度、压强及它们的交互作用对化学物品产率的影响。

```
proc glm data=chap6.outcome;                    /*调用 glm 过程*/
class   Temp Pressure;                          /*定义分类变量为 Temp 和 Pressure*/
model outcome=Temp Pressure Temp*Pressure;
/*定义模型中的因变量为产率 outcome，自变量为 Temp、Pressure 和它们的交互项*/
run;
```

选择 Run|Submit 命令提交程序，以下分析主要输出结果。

表 6-30 为模型方差分析表，由 *F* 检验 *P* 值<0.0001（<0.05）可知模型显著成立。表 6-31 为模型一些拟合参数，模型的拟合优度（R-Square）达到了 0.935943，则用方差分析方法分析此数据为合理的。表 6-32 为效应方差分析表，效应 Temp 对应的 *F* 检验 *P* 值为 0.0008（<0.05），效应 Pressure 及 Temp 和 Pressure 的交互项的 *F* 检验都小于 0.0001，则可判断温度（Temp）、压强（Pressure）和它们的交互项对化学物品的产率有显著的影响。

表 6-30　模型方差分析表

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	77.25555556	9.65694444	49.31	<0.0001
Error	27	5.28750000	0.19583333		
Corrected Total	35	82.54305556			

表 6-31　模型拟合参数

R-Square	Coeff Var	Root MSE	Outcome Mean
0.935943	0.482512	0.442531	91.71389

表 6-32　效应方差分析表（I 型离差平方和）

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Temp	2	3.69055556	1.84527778	9.42	0.0008
Pressure	2	55.71055556	27.85527778	142.24	<0.0001
Temp*Pressure	4	17.85444444	4.46361111	22.79	<0.0001

（2）接下来编程分析控制温度（压强）为一定水平时，不同的压强（温度）对化学物品的产率的影响。

```
proc glm data=chap6.outcome;                    /*调用 glm 过程*/
class   Temp Pressure;                          /*定义分类变量为 Temp 和 Pressure*/
model outcome=Temp Pressure Temp*Pressure;
/*定义模型中的因变量为产率 outcome，自变量为 Temp、Pressure 和它们的交互项*/
lsmeans Temp*Pressure/slice=Temp slice=Pressure;
run;
```

选择 Run|Submit 命令提交程序，以下分析主要输出结果。

表 6-33 为压强和温度的不同组合下该化学物品的产率的均值，观察得到在温度水平为 B、

压强为 215Pa 时，该化学物品的产率最高（94.175%）。

表 6-33　压强和温度不同水平下产率均值

Temp	Pressure（Pa）	Outcome LSMEAN（%）
A	200	90.3500000
A	215	94.1750000
A	230	89.4750000
B	200	91.1250000
B	215	93.4500000
B	230	90.5000000
C	200	92.5250000
C	215	92.5500000
C	230	91.2750000

表 6-34 为控制温度（Temp）比较不同的压强（Temp）下该化学物品的产率，发现无论将温度控制在何种水平之下，生产过程中不同的压强作用与该化学物品的产率有显著的差异。类似的，表 6-35 为控制压强（Pressure）比较不同的温度（Temp）下该化学物品的产率，发现将压强控制在不同水平时生产过程中不同的温度（Temp）对应的该化学物品的产率差异显著。

表 6-34　控制压强比较不同温度下的产率

Temp*Pressure Effect Sliced by Temp for Outcome					
Temp	DF	Sum of Squares	Mean Square	F Value	Pr > F
A	2	49.981667	24.990833	127.61	<0.0001
B	2	19.331667	9.665833	49.36	<0.0001
C	2	4.251667	2.125833	10.86	0.0003

表 6-35　控制温度比较不同压强下的产率

Temp*Pressure Effect Sliced by Pressure for Outcome					
Pressure	DF	Sum of Squares	Mean Square	F Value	Pr > F
200	2	9.721667	4.860833	24.82	<0.0001
215	2	5.301667	2.650833	13.54	<0.0001
230	2	6.521667	3.260833	16.65	<0.0001

（3）编写以下程序尝试应用 Contrast 语句实现控制温度在 A 水平（≤20℃）时对不同的压强对应的化学物品产率进行两两比较。

```
proc glm data=chap6.outcome;                    /*调用 glm 过程*/
class   Temp Pressure;                          /*定义分类变量为 Temp 和 Pressure*/
model outcome=Temp Pressure Temp*Pressure;
/*定义模型中的因变量为产率 outcome，自变量为 Temp、Pressure 和它们的交互项*/
lsmeans Temp*Pressure/slice=Temp;
/*控制温度（Temp）计算不同压强（Pressure）对应的化学物品产率的均值*/
```

```
contrast 'P1 vs P3 in Temp1' Pressure 1 0 -1 Temp*Pressure   1 0 -1;
contrast 'P1 vs P2 in Temp1' Pressure 1 -1 0 Temp*Pressure   1 -1 0;
contrast 'P2 vs P3 in Temp1' Pressure 0 1 -1 Temp*Pressure   0 1 -1;
/*以上三行程序为将温度控制在 20℃时，比较不同压强对应的产率*/
run;
```

选择 Run|Submit 命令提交程序，以下分析两两比较结果（如表 6-36 所示），第二行给出的是控制温度在小于或等于 20℃时，比较压强在 200Pa 和 230Pa 对应的化学产率的差异，由于 *P* 值为 0.0094，则此均值差异显著。即比较温度小于或等于 20℃时，压强在 200Pa 和 230Pa 时生产的化学物品产率的差异显著。类似的，读者自行分析表中其他结果。本例中的 Contrast 语句只为抛砖引玉，希望读者在实际应用中根据具体情况合理使用。

表 6-36　两两比较结果

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
P1 vs P3 in Temp1	1	1.53125000	1.53125000	7.82	0.0094
P1 vs P2 in Temp1	1	29.26125000	29.26125000	149.42	<0.0001
P2 vs P3 in Temp1	1	44.18000000	44.18000000	225.60	<0.0001

综上所述，观察表 6-32 可知当控制温度小于或等于 20℃，压强为 215Pa 该化学物品的产率最高，则此条件相对而言是最优生产条件；且温度和压强都是显著地影响该化学物品的产率，若希望得到更精确的试验结果，寻找更优的产率的生产条件，可考虑对温度和压强因素安排更多的水平。

菜单法：

步骤一：选择 Solutions|Analysis| Analyst 命令，进入 Analyst 分析界面。

步骤二：选择 File|Open|Open By Sas name|chap6|outcome|OK 命令，打开数据集 chap5.outcome。

步骤三：选择 Statistics|ANOVA|Factorial ANOVA 命令，弹出如图 6-10 所示对话框，单击变量 Outcome，再单击 Dependent（因变量）按钮，则定义因变量为 Outcome。类似的，定义自变量（Independent）为 Pressure 和 Temp。

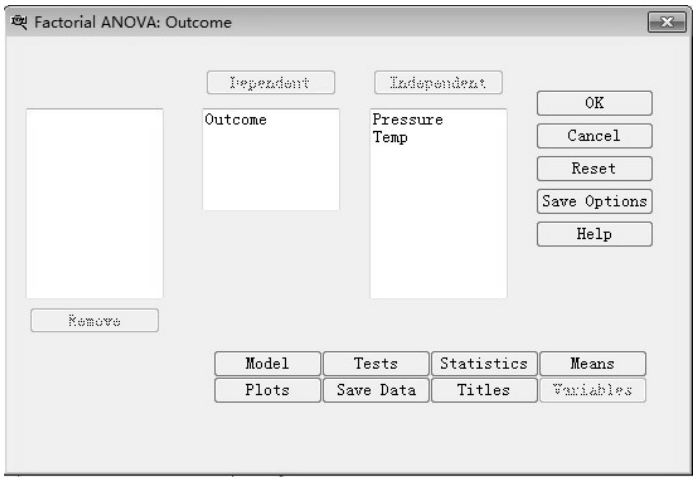


图 6-10　方差分析主对话框（析因分析）

步骤四：单击 Model（模型）按钮，弹出如图 6-11 所示对话框，单击 Standard Models（标准模型）按钮，在出现的下拉菜单中单击 Effects up to 2-way interactions（两因素交互作用项），则确定模型主效应为 Temp、Pressure、Temp*Pressure。单击 OK 按钮保存设置并返回如图 6-10 所示对话框。单击 OK 按钮，则将输出析因设计方差分析结果，注意应用菜单法不能控制某一个（几个）效应两两比较某一效应不同水平下因变量取值的差异。

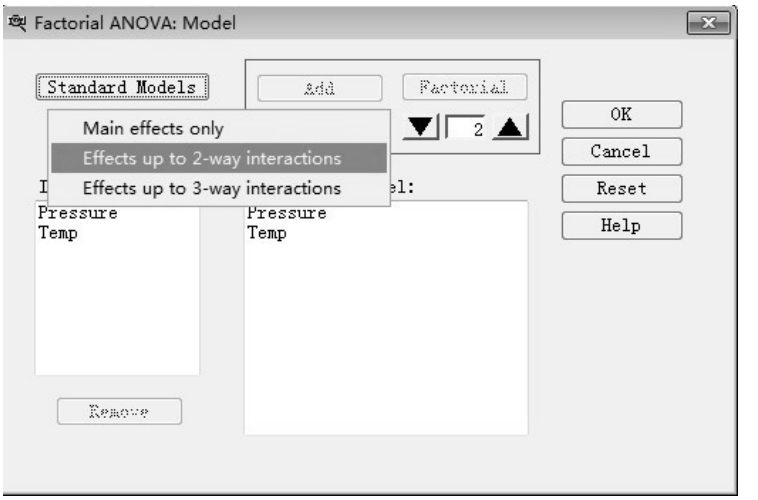


图 6-11　选择分析模型

6.7　协方差分析

6.7.1　基本原理

协方差分析是结合回归与方差分析的一种分析方法。在设计试验研究因变量时，为了将试验误差降到最低，准确地获得处理因素的试验效应，希望其他可能影响和干扰因变量的变量保持一致。若这些干扰变量难以或不能控制时，在试验中同时记录它们的值，并把它们看作或称协变量，建立因变量随着协变量变化的回归方程，将不同组的因变量调整到协变量相等时再进行差异性分析。从而较合理地检验在效应项的不同水平下，修正后的因变量的均值差异是否显著。

进行协方差分析需要满足以下假定条件：

- 各组（按效应水平分组）样本来自具有相同方差 σ^2 的正态分布总体。
- 协变量与因变量之间的总体回归系数不等于 0。
- 各组（按效应水平分组）回归线平行。

对随机区组、析因、拉丁方等实验设计数据进行方差分析时都可以引入一个或多个协变量，扣除协变量的影响后比较因变量的修正均值。

用 SAS 中的 GLM 过程进行协方差分析时，要注意不同试验设计时 class 语句和 model 语句的写法。设分类变量为 A、B，协变量为 X，因变量为 Y，则有：

单因素 k 水平设计的协方差分析模型：

```
class     A;
model     Y=X    A ;
```

随机区组设计的协方差分析模型：

```
class     A   B;
model     Y=X   A   B ;
```

两因素析因设计的协方差分析模型：

```
class     A   B;
model     Y=X   A   B   A*B;
```

6.7.2 SAS 实例——比较不同化肥对桃子的产量的影响

例 6-5 为考量三种不同的化肥对桃子产量的影响，从试验农田中随机选择了 30 棵桃树分成三组，分别记录下施三种不同的化肥前后桃树的产量（kg/棵），数据如表 6-37 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap6\product”）。试比较不同化肥对桃树增产的效果。

表 6-37 桃子产量数据 单元：kg/棵

肥料（Fertilizer）	类别	产量（Product）									
A	施肥前	47	58	53	46	49	56	45	44	49	60
	施肥后	56	68	65	53	58	68	63	52	58	70
B	施肥前	55	56	67	61	62	64	66	69	54	56
	施肥后	60	60	73	68	68	69	70	76	59	6
C	施肥前	44	48	46	50	49	57	58	53	43	47
	施肥后	52	58	54	61	70	64	69	66	50	57

解析：注意到即便是随机挑选的 30 棵桃树，并且随机分组，但桃树的产量和桃树本身的大小可能存在一定的影响，因此本例将分别采用常规的方差分析和将桃树施肥前的产量作为协变量采用协方差分析比较结果。

编程法：

单因素方差分析法，仅仅比较施用不同类别的化肥后桃树的产量。

编写程序如下所示：

```
proc anova data=chap6.product;          /*调用 anova 过程*/
class fert;                             /*定义分类变量 fert*/
model pro_aft=fert;                     /*定义模型因变量为 pro_aft，自变量为 fert*/
means fert/bon;                         /*用 bon 法两两比较不同的化肥作用下桃子的产量*/
run;
```

选择 Run|Submit 命令提交程序，以下分析主要输出结果：对模型（如表 6-38 所示）和效应（如表 6-39 所示）的显著性检验 P 值为 0.0631（>0.05），同时 Bonferroni 法多重检验结果将三种化肥归为一类，联系本实验背景，即施用三种不同的化肥的桃树的产量没有显著的差异，如表 6-40 所示。

表 6-38　模型方差分析表

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	253.066667	126.533333	3.07	0.0631
Error	27	1113.900000	41.255556		
Corrected Total	29	1366.966667			

表 6-39　效应方差分析表

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Fert	2	253.0666667	126.5333333	3.07	0.0631

表 6-40　Bonferroni 多重比较结果

Means with the same letter are not significantly different.			
Bon Grouping	Mean	N	Fert
A	66.700	10	B
A			
A	61.100	10	A
A			
A	60.100	10	C

施用三种化肥的桃树产量果真没有显著的差异吗？编写以下程序，用协方差分析方法将桃树的施肥前的产量作为协变量纳入分析，看结果是否一致。

```
proc glm data=chap6.product;                    /*调用 glm 过程*/
class fert;
model pro_aft=fert pro_bef fert*pro_bef;        /*定义模型*/
run;

proc glm data=chap6.product;
class fert;
model pro_aft=fert pro_bef/solution;            /*定义模型类型和输出结果*/
lsmeans fert/stderr pdiff;    /*输出因变量 pro_aft 的修正均值，并进行校正后两两均值比较*/
run;
```

选择 Run|Submit 命令，以下分析主要输出结果。

（1）第一个 GLM 过程（协方差）分析结果：本过程的目的为考察实例数据是否满足协方差分析的条件“各组回归线平行”。观察交互项 pro_bef*Fert 的检验 *P* 值为 0.8805（>0.05），则交互项在模型中不显著，则可近似认为满足协方差分析的假定条件。观察表 6-41 得到效应项 pro_bef（即施肥前的桃树产量）的检验 *P* 值小于 0.0001，即它对 pro_aft（施肥后的桃树产量）的影响显著。因此在方差分析中不能忽略它的影响，应该将它取作协变量进行协方差分析。

表 6-41　效应项的方差分析表（Type I SS）

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Fert	2	253.0666667	126.5333333	11.81	0.0003
pro_bef	1	853.9982927	853.9982927	79.70	<0.0001
pro_bef*Fert	2	2.7410441	1.3705220	0.13	0.8805

（2）第二个 GLM 过程（协方差）分析结果：观察得到表 6-42 所示对主效应 Fert 和 pro_bef 的显著性检验 *P* 值都小于 0.05，因此桃树施肥前产量和施肥的种类对施肥后的桃树产量影响显著。

表 6-42　效应方差分析表（Type=I）

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Fert	2	253.0666667	126.5333333	12.66	0.0001
pro_bef	1	853.9982927	853.9982927	85.43	<0.0001

表 6-43 为修正因变量（剔除协变量影响后）信息，表中第二、三、四列分别为剔除桃树施肥前产量的影响后施肥后桃树产量的均值、标准误、零均值检验 *P* 值。施 C 种肥料后的桃树产量修正均值（64.4831952）相对最大。

表 6-43　修正因变量

Fert	pro_aft LSMEAN	Standard Error	Pr > |t|	LSMEAN Number
A	64.2407147	1.0559749	<0.0001	1
B	59.1760900	1.2892805	<0.0001	2
C	64.4831952	1.1065744	<0.0001	3

表 6-44 为修正均值两两比较结果。表中纵列和横列代表 Fert 的三个水平，单元格是两两比较的 *P* 值。如化肥 A 和化肥 C 对应的 *P* 值为 0.8658（>0.05），则这两种方法对应的因变量修正均值差异显著。分析此表得出结论：剔除施肥前桃树产量的影响，施 A 种和 B 种化肥、B 种和 C 种化肥的施肥后桃树产量有显著差异。

表 6-44　修正均值两两比较结果

Least Squares Means for effect Fert　Pr > |t| for H0: LSMean(i)=LSMean(j)　Dependent Variable: pro_aft			
i/j	1	2	3
1		0.0101	0.8658
2	0.0101		0.0101
3	0.8658	0.0101	

调用 GLM 过程进行协方差分析和调用 ANOVA 过程进行一般方差分析的结论不一致。因此若方差分析过程中的混杂因素影响显著时应考虑将其作为协变量进行协方差分析，以期得到更精准的结果。

菜单法：

步骤一：选择 Solutions|Analysis| Analyst 命令，进入 Analyst 分析界面。

步骤二：选择 File|Open|Open By Sas name|chap6|product|OK 命令，打开数据集 chap6.product。

步骤三：选择 Statistics|ANOVA|Linear ANOVA 命令，弹出如图 6-12 所示对话框，单击变量 pro_aft，再单击 Dependent（因变量）按钮，定义 pro_aft 为模型因变量，类似的，定义 Fert 为分类变量（Class），定义 pro_bef 为协变量（Quantitative）。

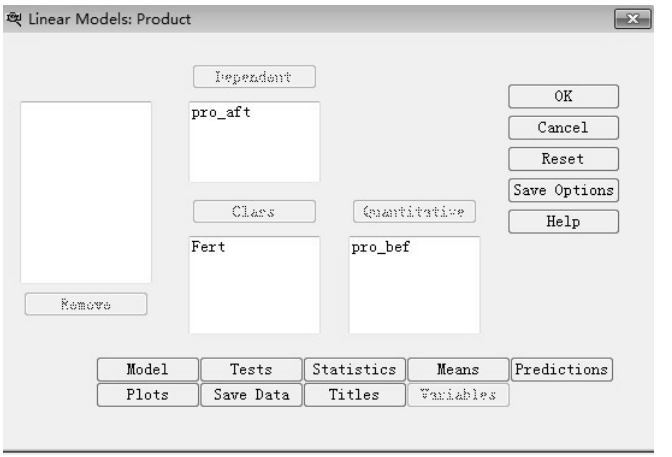


图 6-12　方差分析主对话框

步骤四：单击 Model（模型）按钮，弹出如图 5-21 所示对话框，单击 Standard Models（标准模型）选项，单击下拉菜单中的 Main effect only（仅模型主效应）按钮，则变量 pro_bef 和 Fert 自动被选进模型主效应（Effects in model）选项框。单击 OK 按钮，保存设置并返回如图 6-12 所示对话框。

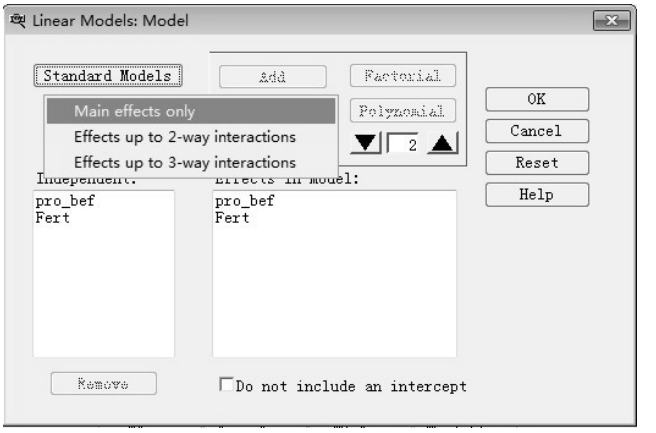


图 6-13　定义模型

步骤五：单击 Means（均值）按钮，在弹出的对话框中单击 LS Means（最小二乘均值），则显示相应的选项卡，如图 6-14 所示。单击变量 Fert，再单击 LS Mean（最小二乘均值）按钮，设定输出扣除施肥前产量的施用不同类肥料以后桃树修正产量均值；勾选 Compute p's for pairwise difference，选项（计算两两均值比较的 *P* 值）。单击 OK 按钮保存设置并返回如图 6-12 所示对话框。单击 OK 按钮，将输出与编程操作相同的结果。

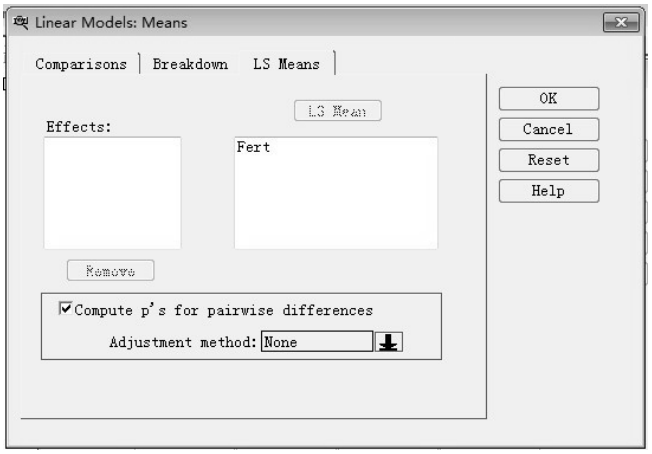


图 6-14　多重比较

练习题

习题 6-1　为了检验不同品牌电池的质量，质检部门抽检了三家生产商生产的五号电池，在每个厂家随机抽取 20 个电池，测得使用寿命（小时）数据如表 6-45 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap6\Battery”）。请问不同品牌电池质量之间是否存在显著差异？（α=0.05）

表 6-45　抽检不同品牌电池寿命数据　　单位：h

试 验 号	电 池 品 牌			试 验 号	电 池 品 牌		
	品 牌 A	品 牌 B	品 牌 C		品 牌 A	品 牌 B	品 牌 C
1	53	47	54	11	56	48	49
2	49	50	55	12	54	48	51
3	50	45	52	13	52	50	50
4	52	45	52	14	51	45	49
5	53	47	52	15	53	47	54
6	50	46	48	16	53	50	49
7	51	47	47	17	54	54	50
8	50	49	54	18	49	49	55
9	53	51	49	19	48	46	55
10	56	52	51	20	52	49	50

（本习题的解答程序在光盘中的存储路径为“proc\chap6\Battery”。）

习题 6-2　研究 7 种硬木浓度，以便确定它们对所生产的纸张强度的效应。但是由于实验装置每天只能进行三次试验，因为日期也可能有影响，实验员用平衡不完全区组设计，如表 6-46 所示（相应的 SAS 数据在光盘中的存储路径为“data\chap6\paper”）。请分析数据并得到结论。

表 6- 46　不同硬木浓度下对纸张效应强度影响的试验数据

硬木浓度（%） （Density）	日期 （Day）						
	一	二	三	四	五	六	七
2	118				122		115
4	126	120				119	
6		137	117				134
8	141		129	149			
10		145		150	143		
12			120		118	123	
14				136		130	128

（本习题的解答程序在光盘中的存储路径为“data\chap6\paper”。）

习题 6-3　比较 A、B、C、D、E 5 种药物给小白鼠注射后产生的皮肤疱疹大小（用疱疹的直径作为试验结果），用 5 只小白鼠进行试验，每只小白鼠有 5 个部位供注射，采用拉丁方试验设计得到数据如表 6-47 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap6\skin”）。请分析不同药物、不同的部位和不同的小白鼠产生的疱疹大小是否有显著的差异。（α=0.05）

表 6-47　小白鼠注射药物后皮肤疱疹记录数据

部位 （Location）	小白鼠（Mice）				
	一	二	三	四	五
甲	A(0.5)	B(0.8)	C(0.5)	D(0.4)	E(0.7)
乙	E(0.6)	A(0.6)	B(0.6)	C(0.3)	D(0.8)
丙	D(0.5)	E(0.9)	A(0.5)	B(0.2)	C(0.9)
丁	C(0.7)	D(0.7)	E(0.7)	A(0.5)	B(1.0)
戊	B(0.8)	C(1.0)	D(0.9)	E(0.7)	A(1.2)

（本习题的解答程序在光盘中的存储路径为“proc\chap6\skin”。）

习题 6-4　一位细菌学家对病毒考察两种不同的培养基的效应和两种不同的滋长时间的效应。她采用有 6 次重复的 2^2 设计，以随机顺序做试验。得到的试验结果如表 6-48 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap6\Virus”）。请分析这两种效应及它们的交互作用对病毒生长时间的影响。

表 6-48　病毒生存数据

时　间 （Time）	培养基（Medium）			
	A		B	
12 小时	21	22	25	26
	23	28	24	25
	20	26	29	27
18 小时	37	39	31	34
	38	38	29	33
	35	36	30	35

（本习题的解答程序在光盘中的存储路径为“proc\chap6\Virus”。）

习题 6-5 检验一工业粘接剂的 4 种不同的配方。粘接剂的抗压强度与厚度有关。对每种配方得出强度（单位：磅）和厚度（单位：cm）的 5 个观察值，数据如表 6-49 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap6\Adhesive”）。请分析不同配方的粘接剂的抗压强度是否存在显著差异？（α=0.05）

表 6-49 不同配方粘接剂的强度和厚度

粘接剂配方（Formula）							
1		2		3		4	
强度 Strength	厚度 Thickness	强度 Strength	厚度 Thickness	强度 Strength	厚度 Thickness	强度 Strength	厚度 Thickness
46.5	13	47.8	12	46.3	15	44.7	16
45.9	14	48.0	10	47.1	14	43.0	15
49.8	12	50.2	10	48.9	11	51.0	10
46.1	12	48.6	11	48.2	11	48.1	12
44.3	14	45.8	12	50.3	10	48.6	11

注：强度单位为磅，厚度单位为 cm。

（本习题的解答在光盘中的存储路径为“proc\chap6\Adhesive”。）

第 7 章　相关与回归分析

相关分析用来考察两个变量的相互变化的关联关系，变量间没有因果关系。回归分析用于研究可测量的变量之间的关系，根据变量间的关系，由一个或几个变量来预测另一个变量的取值，一般的分析步骤为：确定分析变量；构造回归模型；诊断模型；利用模型进行描述、控制和预测。相关和回归分析的主要区别为：相关分析的变量之间地位平等，而回归分析变量之间有一定依存关系。

SAS/BASE 中的 CORR 过程常用来进行相关分析，在 SAS/STAT 中有多个实现回归分析的过程，本章将要介绍的有：

- REG 过程——处理一般回归分析，可诊断及简化模型。
- GLM 过程——一般线性模型，自变量可为类别变量或多项式。
- LOGISTIC 过程——主要用于处理因变量为离散型数据的回归模型。
- NLIN 过程——建立非线性回归模型。

感兴趣的读者可以参看 SAS 软件自带的工作手册，根据数据特征和实际情况选择合适的回归分析过程。

7.1　相关分析

7.1.1 基本原理

本节将介绍两个常用的衡量相关性的指标：Pearson 相关系数与 Spearman 相关系数，主要介绍它们的计算公式和适用类型。

Pearson 相关系数用来衡量数值型变量间的线性关系。如衡量国民收入和居民储蓄存款、身高和体重、语文成绩和数学成绩等变量间的线性相关关系。两个随机变量 X 和 Y 的相关系数的计算公式为：

$$R_{\text{pearson}}=\frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X\sigma_Y}$$

其中，μ_X 和 μ_Y 分别代表变量 X 和 Y 的均值；σ_X 和 σ_Y 分别代表变量 X 和 Y 的标准差。

Spearman 相关系数是衡量变量 X 和变量 Y 之间是否存在等级相关的指标，适用于不能准确地测量指标值而只能以严重程度、名次先后等定序的等级资料或某些不呈正态分布或难于判断分布的资料。

设 R_i 和 Q_i 分别为 x_i 和 y_i 各自在变量 X 和变量 Y 中的秩，Spearman 秩相关系数为：

$$R_{\text{spearman}} = \frac{\text{Cov}(R,Q)}{\sigma_R \sigma_Q} = \frac{E((R-\mu_R)(Q-\mu_Q))}{\sigma_R \sigma_Q}$$

两个相关系数在形式上完全一致。

7.1.2 SAS 过程——CORR 过程

CORR 过程适用数值型随机变量的相关分析，可计算 Pearson 相关系数、Spearman 秩相关系数、Kendall's tau-b 统计量、Hoeffding's 独立性分析统计量 *D* 及 Spearman、Pearson、Kendall 偏相关系数。调用 CORR 过程进行的相关分析系统默认给出分析变量的描述性统计量分析结果、Pearson 相关系数及每个变量对应的零均值检验 *P* 值。

PROC CORR 过程的一般使用格式如下：

```
PROC CORR   DATA=SAS 数据集 <选项>;
VAR         变量列表;
WITH        变量列表;
PARTIAL     变量列表;
WEIGHT      变量;
FREQ        变量;
BY          变量列表;
RUN;
```

PROC CORR 语句是过程中唯一必须定义的语句，其后主要的控制选项如表 7-1 所示。

表 7-1　PROC CORR 语句后主要控制选项

选　项	意　义
OUTP＝SAS 数据集	新建一个包含 Pearson 相关系数的 SAS 数据集
OUTS＝SAS 数据集	新建一个包含 Spearman 等级相关系数的 SAS 数据集
OUTK＝SAS 数据集	新建一个包含 Kendall TB 相关系数的 SAS 数据集
OUTH=SAS 数据集	新建一个包含 Hoeffding *D* 统计量的 SAS 新数据集
PEARSON	计算 Pearson 相关系数（为系统默认）
HOEFFDING	计算并输出 Hoeffding 的 *D* 统计量
KENDALL	计算并输出 Kendall TB 相关系数
SPEARMAN	计算并输出 Spearman 等级相关系数
VARDEF＝DF | WEIGHT|WGT| WDF	指定计算方差时的除数：DF（自由度 N–1）、WEIGHT 或 WGT（权重之和）、N（观察数）、WDF（权重之和−1）。默认值为 DF
COV	计算协方差矩阵
SSCP	要求输出平方和与交叉积和
CSSCP	要求输出离差平方和与交叉积和
NOPRINT	关闭所有打印输出
RANK	要求按绝对值从高到低的次序对每个变量输出相关系数
NOMISS	在计算中去除包含缺失值的观测

续表

选　项	意　义
NOSIMPLE	不输出变量的简单描述性统计量
PLOTS=MATRIX/SCATTER	指定输出散点图，如其取值为 MATRIX，则输出散点图的矩阵；如其取值为 SCATTER，则将输出变量的两两散点图

CORR 过程中使用的语句含义如下：

VAR 语句——定义计算相关系数的变量，否则系统将计算数据集中所有数值型变量的两两相关系数。

WITH 语句——和 VAR 语句联合使用定义计算变量间特殊组合的相关系数。VAR 语句和 WITH 语句列出的变量分别在输出相关矩阵的上方和左边。例如，在程序中定义以下语句：

```
VAR A B;
WITH X Y Z;
```

将生成 X 和 A、Y 和 A、Z 和 A、X 和 B、Y 和 B、Z 和 B 的相关矩阵。

PARTIAL 语句——与 PEARSON、SPEARMAN、KENDALL 等选项一起使用，用来计算净相关系数。目的为在计算 VAR、WITH 语句定义的变量的相应统计量时排除 PARTIAL 语句中定义的变量对它们的值的影响。

WEIGHT 语句——定义加权变量，仅用于计算 Pearson 加权相关系数。

FREQ 语句——指定频数变量，变量值代表观测重复数或加权值的大小。

BY 语句——定义分层变量。

7.1.3　SAS 实例——考察航空公司航班正点率和顾客投诉次数的关系

例 7-1　随机抽取 10 家航空公司，对其最近一年的航班正点率和顾客投诉次数进行了调查，所得数据如表 7-2 所示（相应的 SAS 数据集在光盘中的存储位置为“data\chap7\complaint”）。请问航班正点率和顾客投诉次数之间是否存在相关关系？方向和强度如何？

表 7-2　航班正点率和顾客投诉次数数据

航空公司编号	航班正点率（%）	顾客投诉次数
1	82.8	19
2	76.6	54
3	75.4	73
4	76.2	65
5	73.8	53
6	72.2	92
7	71.2	83
8	70.8	120
9	91.4	15
10	68.2	123

解析：考察两个变量之间的相关关系，可以绘制两者的散点图，观察整体的变化趋势，同时计算相关系数，从数量上得到两者的相关关系的方向和强度。

编写如下程序（其在光盘中的存储路径为“proc\chap7\complaint”）：

```
ods graphics on;
proc corr data=chap7.complaint plots=scatter spearman pearson;
/*调用 corr 过程指定输出散点图，并计算 Pearson 和 Spearman 相关系数*/
var rate number;
run;
ods graphics off;
```

选择 Run|Submit 命令提交程序，以下分析主要输出结果。

如图 7-1 所示散点图以变量 rate（航班正点率）为横坐标，变量 number（顾客投诉次数）为纵坐标。观察此图，得出 number 的值随着 rate 的值的增加呈递减趋势，联系实例背景，随着航班正点率的增加，顾客的投诉次数呈减少趋势。

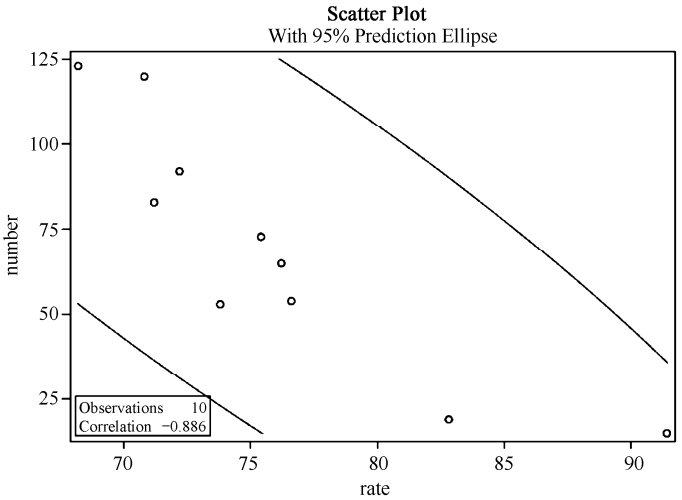


图 7-1　航班正点率和顾客投诉次数散点图

表 7-3 和表 7-4 分别是 Pearson 和 Spearman 相关系数矩阵。变量 rate 和 number 之间的 Pearson 相关系数为−0.88591，Spearman 相关系数为−0.91515，对原假设“该相关系数为零”的检验 *P* 值分别为 0.0006 和 0.0002，均小于 0.05，则两类相关系数显著不为零。联系本例背景，判定航班正点率和顾客投诉次数呈线性组合的负相关关系。

表 7-3　Pearson 相关系数

Pearson Correlation Coefficients, N = 10 Prob > |r| under H0: Rho=0		
	rate	number
rate	1.00000	−0.88591
rate		0.0006
number	−0.88591	1.00000
number	0.0006	

表 7-4　Spearman 相关系数

Spearman Correlation Coefficients, N = 10 Prob > |r| under H0: Rho=0		
	rate	number
rate rate	1.00000	−0.91515 0.0002
number number	−0.91515 0.0002	1.00000

7.2　直线回归

7.2.1　基本原理

首先用含一个自变量的回归模型来简介一元直线回归分析过程，包括模型构建、参数估计、模型检验和模型诊断。

1．模型构建

设基本模型为直线 $Y_t=\alpha+\beta X_t+\varepsilon_t$，其中 Y_t 是第 t 次试验中因变量的取值，α 和 β 是参数，X_t 为第 t 次试验中自变量的取值，ε_t 是满足以下三个条件的随机误差项：

- 均值 $E(\varepsilon_t)=0$；
- 方差 $\mathrm{Var}(\varepsilon_t)=\sigma^2$；
- 当 $i\neq j$ 时，协方差 $\mathrm{Cov}(\varepsilon_i,\varepsilon_j)=0$。即对所有的 $i\neq j$，ε_i 与 ε_j 互不相关。

模型中 Y 的第 t 次观察值 Y_t 包括两部分：常数项 $\alpha+\beta X_t$ 和随机项 ε_t。由于 $E(\varepsilon_t)=0$，因此 $E(Y_t)=\alpha+\beta X_t+E(\varepsilon_t)=\alpha+\beta X_t$，其中 $\alpha+\beta X_t$ 是常数。因此，当第 t 次试验中 X 取为 X_t 时，相应的 Y_t 来自均值是 $E(Y_t)=\alpha+\beta X_t$ 的概率分布，则模型的回归函数是：

$$E(Y)=\alpha+\beta X$$

对任何给定的 X，回归函数把 X 水平与 Y 的概率分布均值联系起来。

在第 t 次试验中，Y 的观察值超过或低于回归函数值的部分为误差项部分 ε_t。假设误差项 ε_t 具有相同的方差 σ^2，则相应的 Y_t 的方差为 $\mathrm{Var}(Y_t)=\sigma^2$，因为 $\mathrm{Var}(Y_t)=\mathrm{Var}(\alpha+\beta X_t+\varepsilon_t)=\mathrm{Var}(\varepsilon_t)=\sigma^2$。无论自变量 X 取值如何，回归总是假设 Y 的概率分布具有相同的方差 σ^2 且假设误差项互不相关。因此任何一次试验的结果对其他各次试验的误差项都没有影响，相应的 Y_i 与 Y_j 也互不相关。

2．估计回归参数

估计回归参数最常用到的是最小二乘法，此方法的基本思想是使所有观测点到直线的“距离”最小，以确定 α 和 β 的值。

设有一组 T 期间内关于两个变量 X 和 Y 的样本观测值（x_t, y_t）（t=1,2,…,N），在 X 和 Y 之间存在函数关系，如果将这些观测数据绘制在图形上，我们会发现所有的点不太可能落在一条直线上，可认为需拟合的回归直线是从分布在平面上的各观测点的中央穿过的直线。以下根据观测数据来确定回归直线的位置（即估计α和β的值）。

确定α和β的值使得所有的观测点和回归直线的“距离”为最小是求解参数的一般规则，假定回归直线为：

$$Y = \alpha^* + \beta^* X$$

则同 $X=X_t$ 对应的估计直线上的点是 $\alpha^* + \beta^* x_t$ 。观测点（x_t,y_t）同估计直线垂直方向的间隔 $e_t = y_t - (\alpha^* + \beta^* x_t)$ 被称为残差。注意误差项和残差的区别：误差项是未知回归直线同观测点的差距，而残差是已知的估计直线同观测点的差距。

最小二乘法参数估计即为求解使以下评价函数取最小值的 α^* 和 β^*：

$$V = \sum_{t=1}^{N} (y_t - \alpha^* - \beta^* x_t)^2$$

以下利用高等数学中寻找函数最大、最小点的方法，令 V 对 α^* 和 β^* 的偏导数为零推导出关于 α^* 和 β^* 的二元联立一次方程组：

$$\begin{cases} \dfrac{\partial V}{\partial \alpha^*} = -2\sum_{t=1}^{N} (y_t - \alpha^* - \beta^* x_t) = 0 \\ \dfrac{\partial V}{\partial \beta^*} = -2\sum_{t=1}^{N} x_t (y_t - \alpha^* - \beta^* x_t) = 0 \end{cases}$$

解联立方程得到参数值如下：

$$\begin{cases} \beta^* = \dfrac{\sum_{t=1}^{N} (x_t - \overline{x})() y_t - \overline{y}}{\sum_{t=1}^{N} (x_t - \overline{x})^2} \\ \alpha^* = \overline{y} - \beta^* \overline{x} \end{cases}$$

其中，$\overline{x} = \dfrac{1}{N}\sum_{t=1}^{N} x_t$，$\overline{y} = \dfrac{1}{N}\sum_{t=1}^{N} y_t$ ，能够证明 α^* 和 β^* 分别是α和β的一致无偏估计量。

3．方程检验

通过以上步骤得到回归方程后，方程是否有意义呢，则需要用到方程检验。

求回归方程的目的是寻找反映 y 随 x 变化的统计规律，如果β=0，即不管 x 如何变化 $E(y)$ 的值不变，此时回归方程无意义。因此回归方程检验问题归结为检验原假设 $H_0 : \beta = 0$，常用 F 检验和 t 检验方法进行检验。

F 检验从观察值的离差平方和分解入手。因变量 $y_1, y_2, \cdots, y_N$ 的差异可以用总离差平方和表示为 $\mathrm{TSS} = \sum_{i=1}^{N} (y_i - \overline{y})^2, df_T = N - 1$。造成这一差异的原因有两个方面：一是由于假设 $\beta = 0$ 不真，即对不同的 x 值， $E(y)$ 随 x 而变化，用回归平方和表示为 $\mathrm{RSS} = \sum_{i=1}^{N} (\hat{y}_i - \overline{y})^2, df_R = 1$；二是由其

他一切随机因素引起的差异，用残差平方和表示为 $\mathrm{ESS}=\sum_{i=1}^{N}(y_i-\hat{y}_i)^2, df_E=N-2$。

而总离差平方和分解成回归平方和与残差平方和：

$$\begin{aligned}\mathrm{TSS}&=\sum(y_i-\overline{y})^2=\sum(y_i-\hat{y}_i+\hat{y}_i-\overline{y})^2\\&=\sum(y_i-\hat{y}_i)^2+\sum(\hat{y}_i-\overline{y})^2\\&=\mathrm{ESS}+\mathrm{RSS}\end{aligned}$$

在 $\beta=0$ 为真，RSS 与 $\mathrm{ESS}/(N-2)$ 都是 σ^2 的无偏估计，采用 F 统计量进行检验：

$$F=\frac{\mathrm{RSS}/\sigma^2/1}{\mathrm{ESS}/\sigma^2/(N-2)}=\frac{\mathrm{RSS}}{\mathrm{ESS}/(N-2)}\sim F(1,N-2)$$

在显著性水平为 α 时，如果 $\dfrac{\mathrm{RSS}}{\mathrm{ESS}/(N-2)}<F_\alpha(1,N-2)$，则拒绝原假设，反之接受原假设。

由于估计的参数的分布为：

$$\hat{\beta}\sim N(\beta,\frac{\sigma^2}{\sum(x_i-\overline{x})^2})$$

$$\hat{\alpha}\sim N\left(\alpha,\sigma^2\left[\frac{1}{N}+\frac{\overline{x}^2}{\sum(x_i-\overline{x})^2}\right]\right)$$

在原假设情形下 $\dfrac{\hat{\beta}}{\sigma/\sqrt{\sum(x_i-\overline{x})^2}}\sim N(0,1)$，但其中 σ 未知，常用 $\hat{\sigma}^2=\mathrm{ESS}/(N-2)$ 去代替，从而在 $\beta=0$ 时有：

$$t=\frac{\hat{\beta}}{\hat{\sigma}/\sqrt{\sum(x_i-\overline{x})^2}}=\frac{\dfrac{\hat{\beta}}{\sigma/\sqrt{\sum(x_i-\overline{x})^2}}}{\sqrt{\dfrac{\mathrm{ESS}}{\sigma^2}/(N-2)}}\sim t(N-2)$$

类似得到原假设 $\alpha=0$ 为真时的 t 统计量：

$$t=\frac{\hat{\alpha}}{\hat{\sigma}\sqrt{1/N+\overline{x}^2/\sum(x_i-\overline{x})^2}}=\frac{\dfrac{\hat{\alpha}}{\sigma\sqrt{1/N+\overline{x}^2/\sum(x_i-\overline{x})^2}}}{\sqrt{\dfrac{\mathrm{ESS}}{\sigma^2}/(N-2)}}\sim t(N-2)$$

4．回归诊断

回归诊断主要用于检验关于回归假设是否成立及模型形式是否正确，否则应用最小二乘法求得的回归方程就缺乏理论依据。这些检验主要探究的问题为：

- 残差是否随机分布、是否为正态性、是否存在异方差。
- 高度相关的自变量是否引起了共线性。
- 模型的函数形式是否错误或在模型中是否缺少重要的自变量。
- 样本数据中是否存在异常值。

主要应用如下方法进行诊断：

残差图分析——残差图是以残差为纵坐标，某一合适的自变量为横坐标的散点图。此分析的基本思想是：在回归模型中假定误差项是均值为零、方差相等、独立的正态分布随机变量。若观察到的数据满足模型假定条件，则残差作为误差的无偏估计应反映误差的基本假设特征。结合常见的残差图总结分析如下：若残差图如图 7-2（a）所示，则误差项满足假设条件。若残差图如图 7-2（b）所示，有一个偏离模型很大的观测点（即异常点），若怀疑异常点是因为数据记录或测量有误，应从数据集中删除异常点再重新拟合回归模型。注意谨慎处理异常点，因为异常点可能代表了某些相当重要的观测点。在 REG 过程中常用 Cook's D 统计量来度量异常点的影响。若残差图如图 7-2（c）和图 7-2（d）所示，残差随 x 的增大而增大或先增后减，则说明不同的观测的残差乃至误差项具有不同的方差变化，此为异方差情况。此时应考虑在回归之前对因变量或自变量进行数据变换，使方差稳定后再拟合模型。若残差图如图 7-2（e）所示，表明模型本身具有非线性趋势，提示用户在模型中是否忽略了若干重要的自变量。若残差图如图 7-2（f）所示，表明模型本身具有线性趋势，说明模型选择错误。

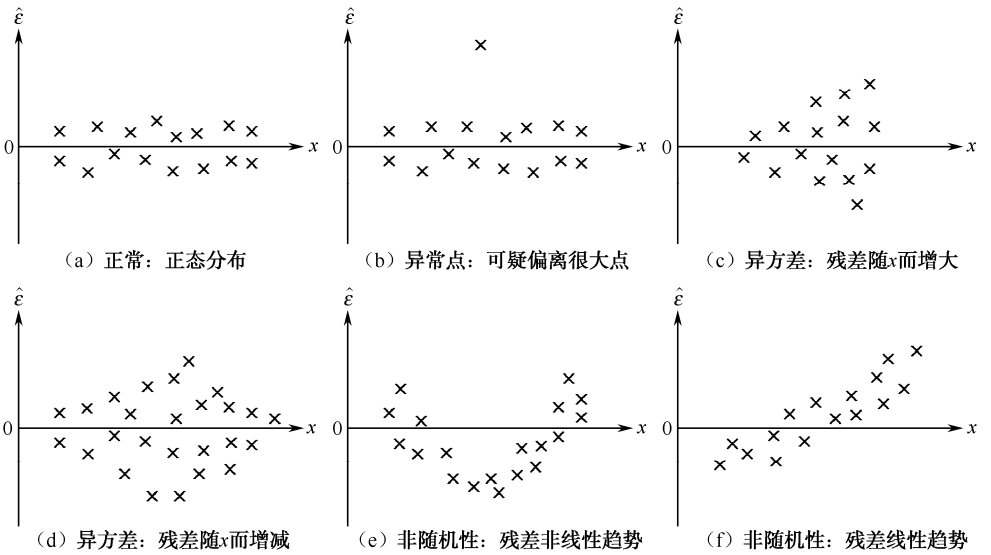


图 7-2　残差图分析

共线性诊断——若回归分析发生模型中两个或两个以上的自变量高度相关，从而引起最小二乘估计可能很不精确。高度相关的自变量及由它们所引起的估计问题合在一起称为共线性问题。共线性诊断问题就是要找出哪些变量间存在共线性关系。REG 过程中提供了特征值法、条件指数 COLLIN 和方差膨胀因子 VIF 统计量进行共线性诊断。

误差项独立性验证——如果误差项不独立，那么我们对回归模型的许多处理，包括误差项估计、假设检验等都将没有推导依据。由于残差是误差的合理估计，因此检验统计量通常是建立在残差的基础上。检验误差独立性的最常用方法，是对残差的一阶自相关性进行 Durbin-Watson 检验。如果 DW 统计量接近于 0，表示残差中存在正自相关；如果 DW 接近于 4，表示残差中存在负自相关；如果 DW 接近于 2，表示残差具有独立性。

7.2.2　SAS 过程——REG 过程

REG 过程是所有回归分析过程中使用最普遍的一种。在过程中用户可应用 MODEL 语句建立用户需要的线性模型，REG 过程提供 9 种选择最佳回归模型的方法，生成数据的散点图和多种统计量；产生部分回归诊断图，并进行共线性诊断；输出预测值、误差、置信区间及向量内成绩矩阵等，并可将这些分析结果存在一个 SAS 数据集中。

REG 过程的一般使用格式如下：

```
PROC REG　DATA=SAS 数据集</选项列表>;
MODEL 因变量=自变量名列 </选项列表> ;
VAR　变量列表;
OUTPUT　OUT=数据集名 </选项列表> ;
PLOT　绘图表达式 </选项列表>;
WEIGHT　变量名;
FREQ　变量名;
BY　变量名;
RESTRICT 方程 1，方程 2，…;
TEST　方程 1，方程 2，…;
RUN;
```

MODEL 语句必须定义，其他语句用户可根据情况选择。REG 过程某些语句后的控制选项如表 7-5 和表 7-6 所示。

表 7-5　PROC REG 语句后主要的控制选项

选　项	意　义
OUTEST=SAS 数据集	输出有关模型的参数估计和选择的统计量到指定 SAS 数据集中
OUTSSCP=SAS 数据集	把平方和及叉积矩阵输出到 TYPE=SSCP 的数据集中
USSCP	输出用在该过程中的所有变量的平方和及叉积矩阵
ALL	输出所有内容
NOPRINT	不输出任何内容

表 7-6　MODEL 语句后主要的控制选项

选　项	意　义
SELECTION =NONE | FORWARD|BACKWARD| STEPWISE | MAXR | MINR | RSQUARE | CP | ADJRSQ	确定变量筛选办法，依次表示全部变量进入法 NONE、前进法 FORWARD、后退法 BACKWARD、逐步筛选法 STEPWISE（前进法与后退法的结合）、最大 R^2 增量法 MAXR、最小 R^2 增量法 MINR、R^2 选择法 RSQUARE、MALLOW'S CP 选择法 CP、修正 R^2 选择法 ADJRSQ
SPEC	进行异方差检验
ACOV	存在异方差时，输出参数β估计量的渐近协方差阵的估计
SLENTRY | SLE =显著性水平	规定变量进入方程的显著性水平
SLSTAY | SLS=剔除水平	规定从方程中剔除变量的显著性水平

续表

选　项	意　义
INCLUDE＝N	迫使前 N 个自变量进入模型
START＝S	从含有 MODEL 语句中前 S 个自变量的模型开始，进行比较、选择过程（仅用于 MAXR 或 MINR 方法）
STOP＝S	当找到最佳的 S 个变量模型之后，逐步回归便停止（仅用于 MAXR 或 MINR 方法）
P	计算各观测点上因变量的预测值
R	作残差分析，并给出因变量的预测值
CLI	计算各自变量对应的因变量的 95%置信区间
CLM	计算各自变量对应的因变量预测值的 95%置信区间
NOINT	指定回归方程不含截距项
STB	输出标准回归系数
COVB	输出回归系数估计的协方差（阵）估计
CORRB	输出回归系数估计的相关矩阵估计
MSE	输出误差项σ^2的估计
RMSE	输出$\hat{\sigma}=\sqrt{\hat{\sigma}^2}$
COLLIN	在未校正截距的情况下，诊断多重共线性，条件数越大越可能存在共线性
COLLINOINT	在校正截距的情况下，诊断多重共线性
TOL	计算共线性水平的容忍度，TOL 值越小说明越可能与别的自变量存在共线性关系
VIF	输出变量间相关性的方差膨胀系数，VIF 越大，说明可能存在共线性，TOL 与 VIF 互为倒数
INFLUENCE	诊断异常点
I	打印$(\boldsymbol{X}'\boldsymbol{X})^{-1}$
XPX	输出模型的$\boldsymbol{X}'\boldsymbol{X}$叉积矩阵
SS1	打印顺序平方和
SS2	打印偏平方和
ALL	输出 SAS 系统分析的以下选择项的特性：XPX，SS1，SS2，STB，COVB，CORRB，SEQB，P，R，CLI，CLM，SPEC，ACOV，TOL, PCORR1，PCOR，R2，SCORR1，SCORR2
PARTIAL	给出每一回归变量的偏回归残差图
DW	计算一阶自相关检验的 DURBIN-WATSON 统计量

INFLUENCE 选项后对异常点的诊断输出的主要统计量如表 7-7 所示。

表 7-7　诊断异常点的统计量

统　计　量	含　义	“异常”的判别准则
LEVERAGE(HI)	杠杆率 HI，第 I 次观测自变量的取值在模型中作用的量度（0≤HI≤1）	HI 越大，则第 I 次观测在模型中的作用就越大
COOK’S D	COOK D 统计量，对某一观测点引起回归影响大小的度量，用于异常点诊断	若 D>50%，则可认为该观测点对模型的拟合有强烈影响
COVRATIO	协方差矩阵的行列式之比（去掉某一观测点后、前对比）	若| COVRATIO | ≥3（自变量个数+I），则第 I 个观测点值得引起注意

OUTPUT 语句：用于把一些计算结果输出到指定的数据集中，其后的关键字及其意义如表 7-8 所示。

表 7-8　OUTPUT 语句后的关键字

关 键 字	意　义	关 键 字	意　义	关 键 字	意　义
PREDICTED	预测值	L95M	95%CLM 下限	STDP	CLM 的标准差
RESIDUAL	残差	U95M	95%CLM 上限	STDR	残差的标准差
PRESS	残差/(1－HI)	L95	95%CLI 下限	STDI	CLI 的标准差
RSTUDENT	刀切残差	U95	95%CLI 上限	COOKED	COOK D 统计量
STUDENT	学生氏残差	H	杠杆点统计量 HI		

注：CLM 代表因变量均值，CLI 代表因变量预测值。

REG 过程中使用的语句含义如下：

VAR 语句——列出叉积矩阵中的变量，仅当定义选项 OUTSSCP=SASDATASET 时使用。

PLOT 语句——绘制两变量的散点图。语句格式为：PLOT　X*Y / 选项。其中 X 和 Y 变量可为原始数据集中的变量或统计量关键字。注意：若变量是统计量关键字时需要在其后加上圆点“．”。

RESTRICT 语句——要求计算条件最小二乘估计，语句定义的方程就是关于回归系数（用自变量表示）的等式，方程与方程间用逗号分隔。如以下语句：

```
MODEL Y=A1 A2 B1 B2;
RESTRICT　A1+A2=1;
```

表示在自变量 A1 和 A2 的系数和为 1 的条件下求回归系数的最小二乘估计。

TEST 语句——要求进行条件显著性检验，其中条件方程是关于回归系数（用自变量表示）的等式，方程与方程间用逗号分隔。TEST 语句一般不与 RESTRICT 语句同用。如以下语句：

```
MODEL Y=A1 A2 B1 B2;
TEST A1+A2=1;
```

表示自变量 A1 和 A2 的系数和为 1 原假设条件下进行 F 检验。

交互式语句——可以直接在 PROC REG 过程中使用表 7-9 列出的交互式语句。

表 7-9　交互式语句

语　句	意　义
ADD 变量名列表	向模型中增加变量
DELETE 变量名列表	删除原拟合模型中的有关变量
REFIT	重新拟合模型
PRINT	输出有关模型的相关信息

7.2.3　SAS 实例——考察沸点和气压的关系

例 7-2　在 19 世纪四五十年代，苏格兰物理学家 James D.Forbes 试图通过水的沸点来估计海拔高度。他研究了气压和沸点的关系，表 7-10 为从他 1857 年的论文中选取了 17 个地方的数

据（相应的 SAS 数据集在光盘中的存储路径为“data\chap7\Forbes”），Forbes 的理论认为，在观测值范围内，沸点和气压值的对数成一条直线，请验证这种假设，并探索气压和沸点之间是如何联系的？这种关系是强是弱？我们能否根据温度预测气压？

表 7-10　气压和沸点实验数据

案　例　号	沸点（华氏度）	气压（英寸汞柱）
1	194.5	20.79
2	194.3	20.79
3	197.9	22.40
4	198.4	22.67
5	199.4	23.15
6	199.9	23.35
7	200.9	23.89
8	201.1	23.99
9	201.4	24.02
10	201.3	24.01
11	203.6	25.14
12	204.6	26.57
13	209.5	28.49
14	208.6	27.76
15	210.7	29.04
16	211.9	29.88
17	212.2	30.06

解析：本例的目的为建立回归模型实现预测。为典型的一元回归情形，以下调用 REG 过程编程实现分析，并介绍菜单操作法。

编程法：

编写程序如下所示（其在光盘中的存储路径为“data\chap7\Forbes）”：

```
/*数据预处理*/
data chap7.Forbes;
set chap7.Forbes;
log_pre=100*log(Pressure);
run;

ods graphics on;
/*调用 corr 过程进行相关分析*/
proc corr data=chap7.Forbes plots=scattor;
var Temp Log_pre;
run;
/*调用 reg 过程进行回归分析*/
proc reg data=chap7.Forbes;                    /*调用 reg 过程*/
model Log_pre=Temp/r dw;                       /*定义分析模型，且进行残差分析*/
```

```
run;
ods graphics off;
```

注意到 Forbes 的实验假设是“沸点和气压值的对数成一条直线”，因此首先对气压值取对数，但是由于取对数以后的气压值在 3 左右，为了避免研究太小的数，因此我们将气压对数值乘以 100，将数据扩大或缩小若干倍，对研究两个变量之间的关系不构成影响。

选择 Run|Submit 命令提交程序，以下分析主要的输出结果。

1．相关分析结果

观察图 7-3 所示温度与气压对数值的散点图及表 7-11 所示的两个变量的相关系数矩阵，可初步判定沸点和气压的对数值呈现显著的正相关关系（Pearson 相关系数为 0.99748，对应的“原假设为该系数值为 0”的检验 *P* 值小于 0.0001）。

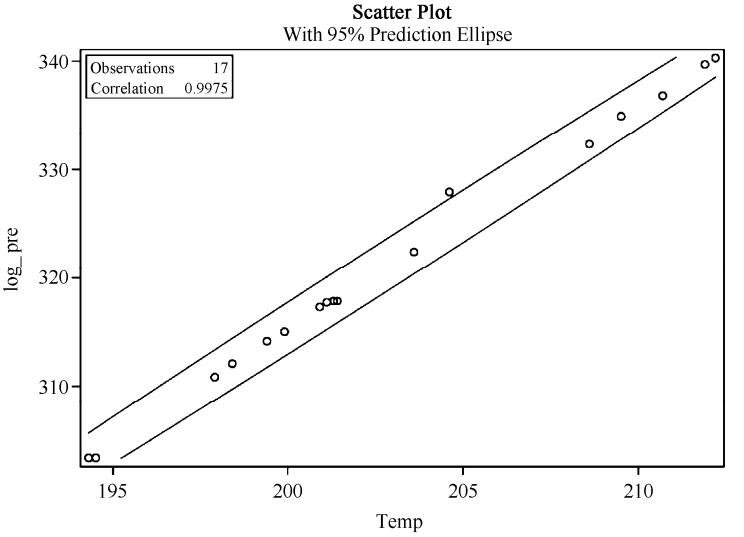


图 7-3　温度与气压对数值散点图

表 7-11　Pearson 相关系数矩阵

Pearson Correlation Coefficients, N = 17 Prob > |r| under H0: Rho=0		
	Temp	log_pre
Temp	1.00000	0.99748
Temp		<0.0001
log_pre	0.99748	1.00000
	<0.0001	

由相关分析初步得到结论：沸点和气压的对数值呈现显著的正相关关系，则用沸点来预测某地的气压值是可行的。

2．直线回归分析结果

模型显著性 *F* 检验的 *P* 值小于 0.001（<0.05），则判断模型有显著意义（如表 7-12 所示）。

表 7-13 为模型拟合统计量：均方根（Root MSE）、因变量均值（Dependent Mean）、变异系数（Coeff Var）、拟合优度（R-Square）、校正拟合优度（Adj R-Sq），其中拟合优度为 0.9950，校正拟合优度为 0.9946，表明模型的拟合度较高。

表 7-12　模型的方差检验

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2257.32063	2257.32063	2961.55	<0.0001
Error	15	11.43315	0.76221		
Corrected Total	16	2268.75378			

表 7-13　模型拟合信息

Root MSE	0.87305	R-Square	0.9950
Dependent Mean	321.45026	Adj R-Sq	0.9946
Coeff Var	0.27160		

表 7-14 为模型的参数估计情况，截距和自变量的参数估计分别为−97.08662 和 2.06224，且对原假设为“此回归系数为零”的 *t* 检验的 *P* 值均小于 0.0001，则回归模型中这两个回归系数都显著不为零。

表 7-14　参数估计与检验

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	−97.08662	7.69377	−12.62	<0.0001
Temp	Temp	1	2.06224	0.03789	54.42	<0.0001

表 7-15 为检验误差项的自相关性结果，DW 统计量为 2.022，在 2 附近，则满足误差项不相关的假定条件。

表 7-15　DW 检验结果

Durbin-Watson D	2.022
Number of Observations	17
1st Order Autocorrelation	−0.027

表 7-16 为残差分析的结果，此表格从左至右每列依次为观测数（Obs）、因变量的观测值（Dependent Variable）、因变量的预测值（Predicted Value）、预测均值的标准误（Std Error Mean Predit）、残差值（Residual）、残差标准误（Std Error Residual）、学生化残差值（Student Residual）、学生化残差图、Cook’s D 统计量。观察学生化残差图，图上出现 4 个及以上的“*”号的那些点代表所对应的学生化残差图绝对值大于 2，被认为是残差较大的可疑点，本例中第

12 个观测为残差过大可疑点，同时观察到它的 Cook's D 统计量为 0.469，也是所有观测中最大的，Cook's D 统计量用来度量因变量对每个观测值对预测值的影响大小，该值越大，则表示观测值对预测值的影响就越大，以此发现强影响点。

表 7-16　残差分析表

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predit	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D
1	303.4472	304.0183	0.3840	−0.5711	0.784	−0.728	| *| |	0.064
2	303.4472	303.6059	0.3903	−0.1586	0.781	−0.203	| | |	0.005
3	310.9061	311.0299	0.2855	−0.1238	0.825	−0.150	| | |	0.001
4	312.1042	312.0610	0.2731	0.0432	0.829	0.0521	| | |	0.000
5	314.1995	314.1233	0.2509	0.0762	0.836	0.0911	| | |	0.000
6	315.0597	315.1544	0.2413	−0.0947	0.839	−0.113	| | |	0.001
7	317.3460	317.2166	0.2256	0.1294	0.843	0.153	| | |	0.001
8	317.7637	317.6291	0.2231	0.1346	0.844	0.160	| | |	0.001
9	317.8887	318.2477	0.2198	−0.3591	0.845	−0.425	| | |	0.006
10	317.8470	318.0415	0.2208	−0.1945	0.845	−0.230	| | |	0.002
11	322.4460	322.7847	0.2132	−0.3386	0.847	−0.400	| | |	0.005
12	327.9783	324.8469	0.2208	3.1314	0.845	3.707	| |******|	0.469
13	334.9553	334.9518	0.3262	0.003470	0.810	0.00428	| | |	0.000
14	332.3596	333.0958	0.3010	−0.7362	0.820	−0.898	| *| |	0.054
15	336.8674	337.4265	0.3620	−0.5591	0.794	−0.704	| *| |	0.051
16	339.7189	339.9012	0.3997	−0.1823	0.776	−0.235	| | |	0.007
17	340.3195	340.5199	0.4094	−0.2003	0.771	−0.260	| | |	0.010

观察残差分析图（如图 7-4 所示）可直观得到表 7-16 所示结果。左上角第一个图的横轴代表因变量预测值（Predicted Value）、纵轴代表残差（Residual），观察此图形状可知残差随着预测值的增加是呈随机分布的。再观察左下角的关于残差的直方图，此图的横轴代表残差（Residual），纵轴代表比率（Percent），可形象得到残差的分布满足近似正态分布。观察左上第二个学生化残差图，此图的横轴代表因变量预测值（Predicted Value）、纵轴代表学生化残差（RStudent），观察此图发现一个观测点的值大于 10，暗示原数据中出现了一个异常值。右下第二个图 Cook's D 统计量图显示第 12 个观测的 Cook's D 统计量的值显著大于其他观测的值，这和表格分析结果一致。若读者在实际应用中发现此类情况，应该考虑重复试验或根据专业知识删除强影响点以获得更精准的结果。

综上分析，建立回归模型为 100log（Pressure）=−97.08662+2.06224Temp。又可写作：$\text{Pressure}=\frac{e^{-97.08662+2.06224\text{Temp}}}{100}$，则建立一个回归预测模型，根据某地的沸点值推测出当地的气压值。

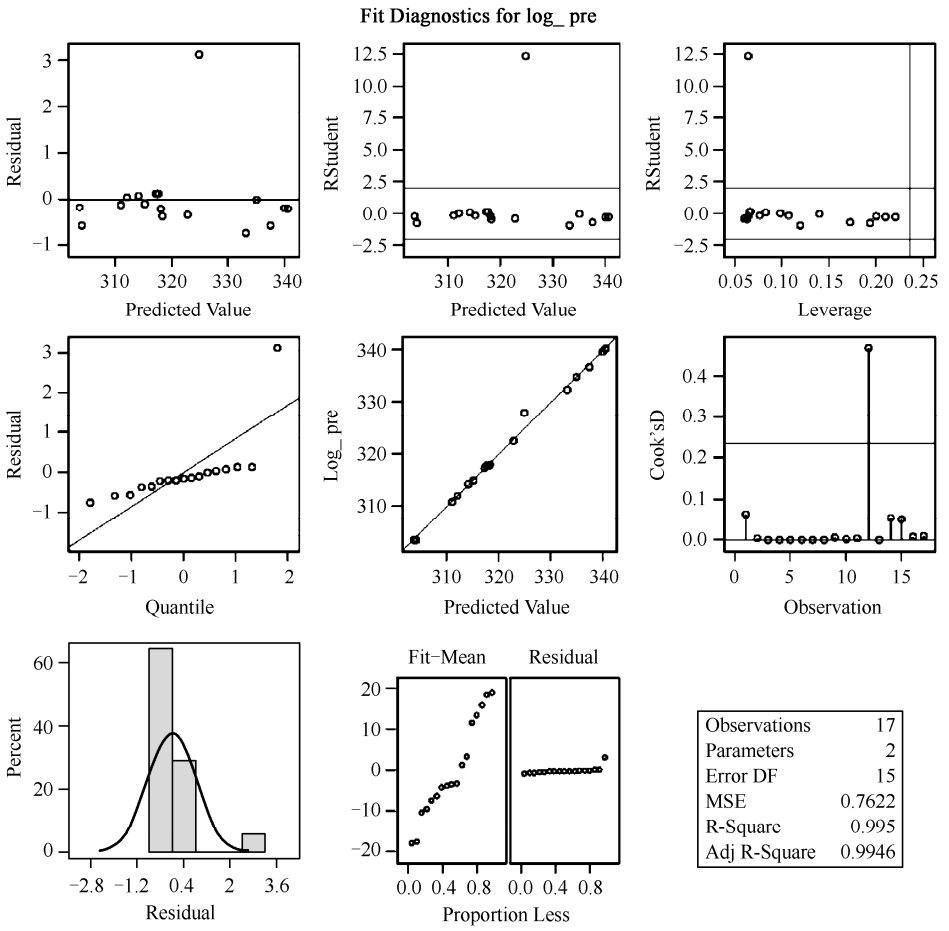


图 7-4　残差诊断图

菜单法：

步骤一：选择 Solutions|Analysis|Analyst 命令，进入 Analyst 分析主界面。

步骤二：选择 File|Open By SAS Name|chap7|Forbes|OK 命令，打开数据集 chap7.Forbes。

步骤三：选择 Statistics|Regression|Simple 命令，弹出如图 7-5 所示对话框，单击变量 log_pre，再单击 Dependent（因变量）按钮，则将变量 log_pre 选为回归模型的因变量。用同样的方式将变量 Temp 选为自变量（Explanatory）。在 Model（模型）选项框内有三个选择：Linear（直线回归模型）、Quadratic（二次抛物线模型）、Cubic（立方抛物线模型），在本例中采用系统默认的 Linear 选项建立直线回归模型。

单击 Statistics（统计量）按钮，弹出如图 7-6 所示对话框，勾选 Parameter estimates（参数估计）下的选项 Std.regression coefficients（标准回归系数）、Confidence limits for estimates（参数的置信区间）、Correlation matrix of estimates（估计值的相关矩阵）、Covariance matrix of estimates（估计值的协方差矩阵），即设置输出以上统计量。单击 OK 按钮保存设置并返回如图 7-5 所示对话框。

单击 Predictions（预测值）按钮，弹出如图 7-7 所示对话框，在 Prediction input（预测值输入）选项框内有以下选项：Predict original sample（预测原始样本）、Predict additional data（预测加选的数据），并在 Data set name（数据集名称）选项框内定义加选的数据集。在 Prediction output

（预测值输出）选项框内有如下选项：List predictions（列出预测值清单）、Save predictions（保存预测值）、Add residuals（添加残差）、Add prediction limits（添加预测限值）。在本例中，勾选 Predict original sample 和 List predictions，即设置对原始值进行预测，并且以清单的形式列出预测值。单击 OK 按钮保存设置并返回如图 7-5 所示对话框。

单击 Plots（绘图）按钮，弹出如图 7-9 所示对话框，在 Predicted（预测图）选项卡内的 Scatter plots（散点图）选项框内可进行如下设置：Plot observed vs predicted（绘制观测值和预测值的散点图）、Plot observed vs independent（绘制观测值和自变量的散点图），还有三个关于图形绘制细节的选项：None（不选）、Confidence limits（加置信区间）、Prediction limits（加预测区间）。在本例中选择 Plot observed vs predicted 和 Prediction limits，即绘制观测值和预测值的散点图，并且加上置信区间。

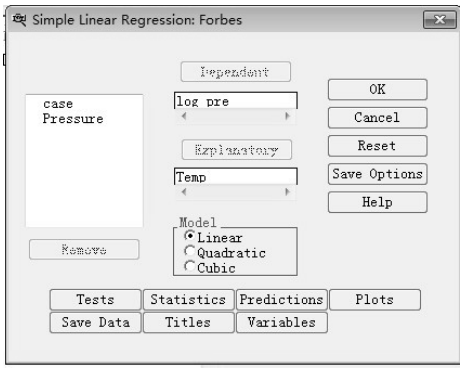


图 7-5　简单回归分析主对话框

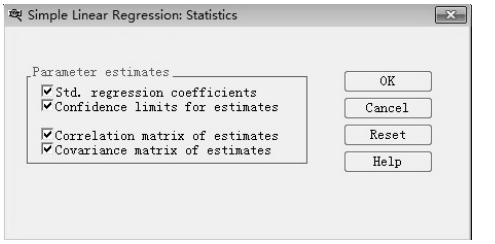


图 7-6　参数估计

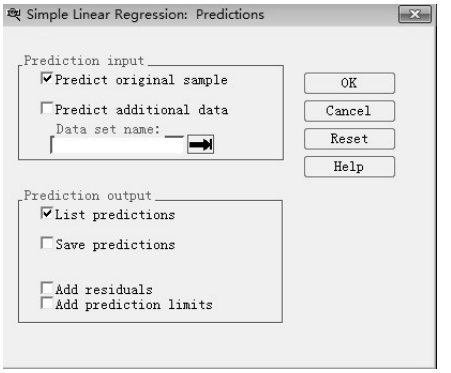


图 7-7　设置预测值

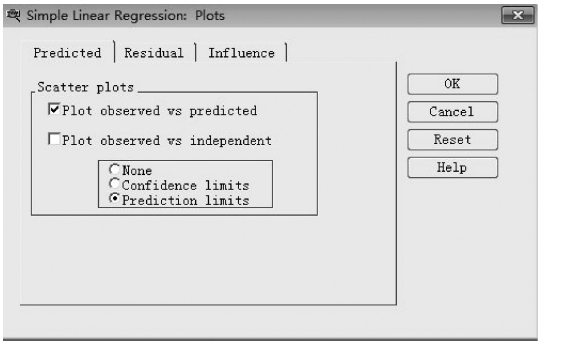


图 7-8　图形绘制（预测图）

单击 Residual（残差图）标签，则显示如图 7-9 所示选项卡。在此设置 Plot residuals vs variables（绘制残差与变量散点图），其中 Residuals（残差）选项框内有三个选项：Ordinary（普通残差）、Standardized（标准化残差）、Studentized（学生化残差）；在 Variables（变量）选项框内有三个选项：Predicted Y（因变量预测值）、Independents（自变量）。勾选 Plot residuals vs variables、Studentized、Independents，即绘制残差与自变量的散点图。若需要绘制正态概率图和分位数图，可勾选 Normal probability-probability plot（正态概率图，即 PP 图）和 Normal quantile-quantile plot（正态概率分位数图，即 Q-Q 图）。

单击 Influence（影响）标签，显示如图 7-10 所示选项卡。在 Influence plots（影响散点图）

选项框内可选择 Plot influence statistics vs variables（绘制统计量与变量的影响散点图），在 Influence statistics（影响统计量）中可选择：DFFITS（由排除特定观察值所引起预测值的变化）、Leverage（H）（杠杆值）、Covariance ratio（协方差比）；在 Variables（变量）选项框内可选择：Predicted Y（因变量预测值）、Independents（自变量）。本例勾选以上所有项目，单击 OK 按钮保存设置并返回如图 7-5 所示对话框。

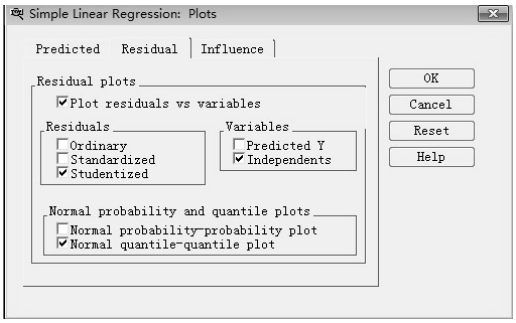


图 7-9　图形绘制（残差图）

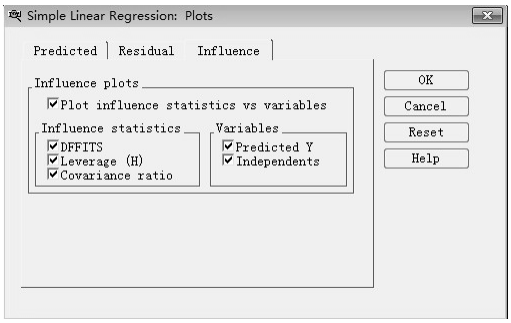


图 7-10　图形绘制（影响图）

单击 Save Data（保存数据）按钮，弹出如图 7-11 所示对话框。勾选 Create and save diagnostics data（建立并保存回归诊断数据），即可保存残差诊断分析的统计量，单击对话框左侧的选项，单击 Add（加入）按钮则可新建包含所选统计量的 SAS 数据集。本例不设置此项目。单击 Cancel 按钮返回如图 7-5 所示对话框。

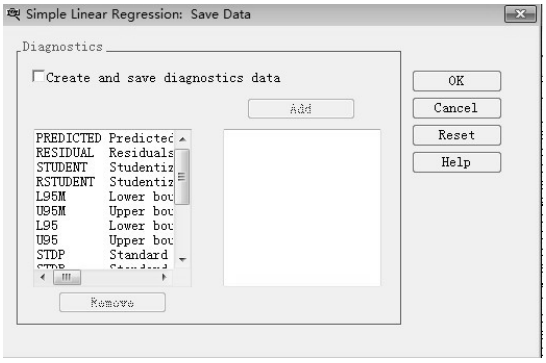


图 7-11　保存统计量

单击图 7-5 所示对话框上的 OK 按钮，则将在 Analyst 结果输出窗口显示结果目录树。单击相应选项即可打开结果。结果中包含回归分析的结果，估计了参数值、参数的协方差和相关矩阵的值，因变量的预测值、自变量和因变量的散点图等。虽然形式和编程所得到的结果不完全一致，但是可得到一致的分析结论。因结果输出过多，在此不一一列举，读者可根据以上步骤自行操作并分析结论。

7.2.4　SAS 实例——多元回归模型预测房屋售价

例 7-3　一幢房子的财产税依赖于目前市场上的房价。由于房子实际的销售较少，当财产税设定时，每栋房子的销售价格每年必须进行估计。回归函数有时候用于得到一个预测函数。

表 7-17 所示数据为宾夕法尼亚的伊利市，*n*=27 幢已经被售出的房子的数据（Nnarula 和 Wellington，1977）相应的 SAS 数据集在光盘中的存储位置为“data\chap7\house”。变量为：

X1=当前税（地方、学校及郡）÷100（美元）

X2=浴室大小

X3=空地大小÷1000（平方英尺）

X4=起居室大小÷1000（平方英尺）

X5=车库数

X6=房间数

X7=卧室数

X8=房子年龄（年）

X9=壁炉数

Y=实际销售价格÷1000（美元）

利用数据估计一个函数，用于由这些 X 及它们的函数预测 Y。

表 7-17　房价数据

X1	X2	X3	X4	X5	X6	X7	X8	X9	Y
4.9176	1.0	3.4720	0.9980	1.0	7	4	42	0	25.9
5.0208	1.0	3.5310	1.5000	2.0	7	4	62	0	29.5
4.5429	1.0	2.2750	1.1750	1.0	6	3	40	0	27.9
4.5573	1.0	4.0500	1.2320	1.0	6	3	54	0	25.9
5.0597	1.0	4.4550	1.1210	1.0	6	3	42	0	29.9
3.8910	1.0	4.4550	0.988	1.0	6	3	56	0	29.9
5.8980	1.0	5.85	1.2400	1.0	7	3	51	1	30.9
5.6039	1.0	9.2500	1.5010	0	6	3	32	0	28.9
15.4202	2.5	9.800	3.4200	2.0	10	5	42	1	84.9
14.4598	2.5	12.8	3.0000	2.0	9	5	14	1	84.9
5.8282	1.0	6.4350	1.2250	2.0	6	3	32	0	35.9
5.3003	1.0	4.9800	1.5520	1.0	6	3	30	0	31.5
6.2712	1.0	5.5200	0.9750	1.0	5	2	30	0	31.0
5.9592	1.0	6.6660	1.1210	2.0	6	3	32	0	30.9
5.0500	1.0	5.0000	1.0200	0	5	2	46	1	30.0
8.2464	1.5	5.1500	1.6640	2.0	8	4	50	0	36.9
6.6969	1.5	6.9020	1.4880	1.5	7	3	22	1	41.9
7.7841	1.5	6.9020	1.4880	1.5	6	3	17	0	40.5
9.0384	1.0	7.8000	1.5000	1.5	7	3	23	0	40.5
5.9894	1.0	5.5200	1.2560	2.0	6	3	40	1	37.5
7.5422	1.5	4.0000	1.6900	1.0	6	3	22	0	37.9
8.7951	1.5	9.8900	1.8200	2.0	8	4	50	1	44.5
6.0931	1.5	6.7265	1.6520	1.0	6	3	44	0	37.9

续表

X1	X2	X3	X4	X5	X6	X7	X8	X9	Y
8.3607	1.5	9.1500	1.7770	2.0	7	3	3	0	38.9
8.1400	1.0	8.0000	1.5040	2.0	7	3	3	0	36.9
9.1416	1.5	7.3262	1.8310	1.5	8	4	31	0	45.8
12.0000	1.5	5.0000	1.200	2.0	6	3	30	1	41.0

解析：本例子属于多元回归分析情形，目的为根据多个指标建立一个回归预测模型。将应用逐步回归的方式选择最终保存在模型中的变量。在实际操作中，读者可以根据数据特征和题目要求选择合适的变量筛选方式。

编程法：

编写程序如下所示（其在光盘中的存储路径为“data\chap7\house”）：

```
proc reg data=chap7.house;    /*调用 reg 过程*/
model Y=X1-X9/selection=stepwise slstay=0.1 slentry=0.15;
/*定义分析模型，并指定逐步回归法进行模型选择*/
run;
```

选择 Run|Submit 命令提交程序，以下按使用的模型选择方法分类分析系统输出的主要结果。SAS 系统将输出每一步模型选择过程中进入模型的变量、模型的显著性检验及进入模型的变量的显著性 *t* 检验。本例省略前 4 步模型选择，直接列出最终（第五步选择后）模型选择结果（如表 7-18 和表 7-19 所示）。

表 7-18　模型方差分析

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5146.22673	1715.40891	118.67	<0.0001
Error	23	332.46068	14.45481		
Corrected Total	26	5478.68741			

表 7-19　模型参数估计

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.97486	2.21266	2.80588	0.19	0.6636
X1	1.78877	0.50220	183.38284	12.69	0.0017
X4	15.36207	2.44066	572.65923	39.62	<0.0001
X9	4.12364	1.82329	73.93713	5.12	0.0335

表 7-20 为此模型每一步选择的描述性统计量，给出了每一步进入的变量、模型总变量个数、进入的自变量对模型的贡献（Partial R-Square）、此模型拟合优度（Model R-Square）、C（P）值、*F* 统计量的值和自变量对应的 *P* 值，观察发现最终模型中包含了 X1、X4、X9 三个统计量，对应的判别系数为 0.9393，说明模型拟合度高。

表 7-20　模型选择中间信息

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	X4		X4	1	0.8685	0.8685	22.2407	165.14	<0.0001
2	X1		X1	2	0.0573	0.9258	4.5236	18.54	0.0002
3	X9		X9	3	0.0135	0.9393	1.8800	5.12	0.0335
4	X2		X2	4	0.0059	0.9452	1.8526	2.37	0.1383
5		X2	X2	3	0.0059	0.9393	1.8800	2.37	0.1383

但是注意到表 7-19 中第二行方程截距（Intercept）对应的 *t* 检验 *P* 值为 0.6634，即不拒绝“该回归方程截距为 0”的原假设。因此编写程序如下所示，拟合去掉截距的回归方程：

```
proc reg data=chap7.house;    /*调用 reg 过程*/;
model Y=X1 X4 X9/noint;
/*定义分析模型，拟合去掉截距的回归模型*/
run;
```

选择 Run|Submit 命令提交程序，以下分析主要输出结果。表 7-21 为模型的方差分析表，由此可知该回归模型对应的 *F* 检验 *P* 值小于 0.0001，则模型为显著成立的。表 7-22 为模型的拟合指数，其中判别系数 R-Square 为 0.9926，校正后的判别系数 Adj R-Sq 为 0.9917,，说明该模型拟合度高。

表 7-21　模型方差分析表

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	45056	15019	1075.12	<0.0001
Error	24	335.26655	13.96944		
Uncorrected Total	27	45392			

表 7-22　模型拟合指数

Root MSE	3.73757	R-Square	0.9926
Dependent Mean	38.44815	Adj R-Sq	0.9917
Coeff Var	9.72107		

表 7-23 为模型的参数估计值及对应的 *t* 检验结果，由此可知变量 X1、X4 和 X9 前对应的系数分别为 1.83293、15.75035 和 3.98724，且都显著不为零（对应的 *t* 检验 *P* 值小于 0.05）。据此可得到回归预测模型：

$$Y = 1.83293X1 + 15.75035X4 + 3.98724X9$$

表 7-23　参数估计

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
X1	X1	1	1.83293	0.48377	3.79	0.0009
X4	X4	1	15.75035	2.23746	7.04	<0.0001
X9	X9	1	3.98724	1.76639	2.26	0.0334

菜单法：

步骤一：选择 Solutions|Analysis|Analyst 命令，进入 Analyst 分析模块。

步骤二：选择 File|Open By SAS Name|chap7|house|OK 命令，打开数据集 chap7.house。

步骤三：选择 Statistics|Regression|Linear 命令，弹出如图 7-12 所示对话框。单击变量 Y，再单击选项 Dependent（因变量），则将定义变量 Y 为回归模型的因变量。类似的，将变量 X1 至 X9 选为自变量（Explanatory）。

单击 Model（模型）按钮，弹出如图 7-13 所示对话框。其中 Model 选项卡中 Selection method（选择方法）选项框内提供了 9 种模型选择方法，选择 Stepwise selection（逐步选择）选项。勾选 Do not include an intercept（不包括截距项）选项，拟合不包括截距项的回归模型。

注意：本例通过编程法，我们得到先验知识：截距项在回归模型中不显著，即不能拒绝“截距项为零”这一原假设。在实际操作中，读者应该反复验证与筛选，根据结果决定是否将截距项保留在模型中。

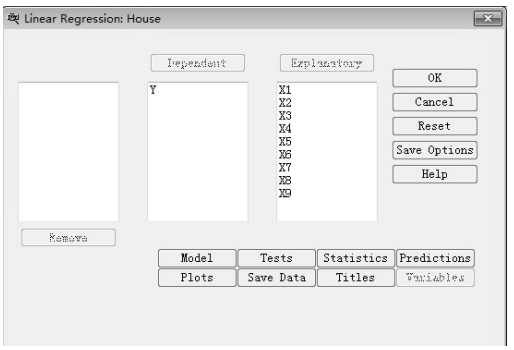


图 7-12　回归分析对话框

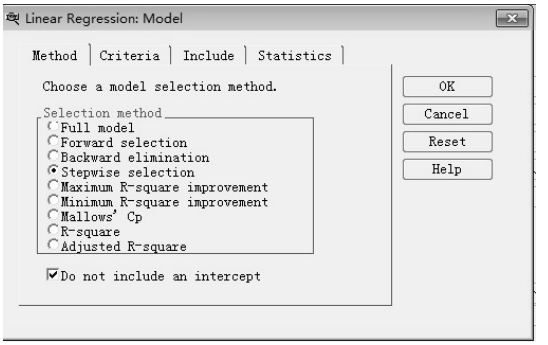


图 7-13　模型选择方法

以下详细介绍图 7-13 所示对话框中 9 种模型选择方法的意义：

Full model（全回归模型）：不筛选回归变量，建立因变量与全部自变量的回归模型。

Forward selection（前向选择法）：选择之初模型中没有变量，首先对每个自变量计算反映其对模型贡献的 F 统计量，将其对应的 P 值与 MODEL 语句中定义的变量被选入模型的显著性水平 α 相比较（SLENTRY=α），若 $P<\alpha$，则此变量被选入模型。Forward selection 将具有最大 F 统计量的变量选入模型，再计算模型外的变量的 F 统计量，重复筛选过程，若再没有比 α 小的 P 值，则 Forward selection 停止。入选模型的变量不会被剔除。系统默认 SLENTRY=0.5。

Backward elimination（向后淘汰法）：选择之初模型中含有全部自变量，依次剔除对模型贡献最小的变量，直到保留对模型中的所有自变量分别计算其 F 统计量，其对应的 P 值都小于 SLSTAY 语句定义的显著性水平。系统默认 STSTAY=0.1。

Stepwise selection（逐步筛选法）：为前向选择的修正，不同之处为逐步筛选法可删除已被选进模型中的变量。在按照前向选择方法选入变量后，再剔除使得 F 统计量的显著性水平低于 SLSTAY 语句定义的水平的变量。在完成检验和必要的剔除之后，其他变量才可再进入模型。当模型外没有变量使 F 统计量的显著性水平在 SLENTRY 语句定义的水平之上，且模型中的每个变量都在 SLSTAY 语句定义的水平之上时，逐步筛选过程结束。系统默认的 SLENTRY=0.15，SLSTAY=0.15。

MAXR（最大 R^2 增量法）：此方法试着找出最佳的一个变量模型、最佳的两变量模型等。MAXR 方法先找出一个产生最大 R^2 值的变量，然后再加入另一个次最大 R^2 值的变量，从而形成二变量的模型，再将模型中的变量与模型外的变量相比较，以决定是否用能生成更大 R^2 值的模型外变量来替换模型中变量，全部比较结束后，便得到了最佳二变量模型。依次往下，得到最佳三变量模型等。

MINR（最小 R^2 增量法）：此方法类似于 MAXR，替换产生最小 R^2 增量的变量。对模型中一个已知的变量数，MAXR 和 MINR 通常产生同样的“最佳”模型。

R-Square（R^2 选择法）：用于寻找某些变量的子集，用户可以规定出现在子集中的自变量的最大和最小个数及被选择的每种子集的个数。此方法可以有效地计算所有可能回归的子集并在每种子集里按 R^2 的递减次序输出这些模型。

ADJRSQ（修正 R^2 选择法）：该方法类似于 R-Square 法，只是对于选择模型使用的准则为修正 R^2 统计量。

Mallows'Cp 统计量：根据 Mallows'Cp 统计量，从模型子集中选出最优子集，通常此值与模型中的变量个数 K 越接近，模型效果越好。

单击 Criteria（标准）标签，显示如图 7-14 所示选项卡。此类选择只适用于前向选择法、后向淘汰法和逐步回归法。在 To enter the model（进入模型）选项中可设置变量进入模型的显著性水平；在 To stay in the model（保存在模型中）选项中可设置变量被剔除模型的显著性水平。可直接输入 0～1 之间的值，或者单击按钮▼或▲减小或增大显著性水平的值，单击一次按钮数值变动幅度为 0.01。本例设置变量进入模型的显著性水平为 0.15，变量被剔除模型的显著性水平为 0.1。

单击 Include（包含）标签，显示如图 7-15 所示选项卡，单击 Variables（自变量）框内的变量名，再单击 Include（包含）按钮，则无论该变量在模型中是否显著，都可强制将其包括在模型中。本例不设置此选项。

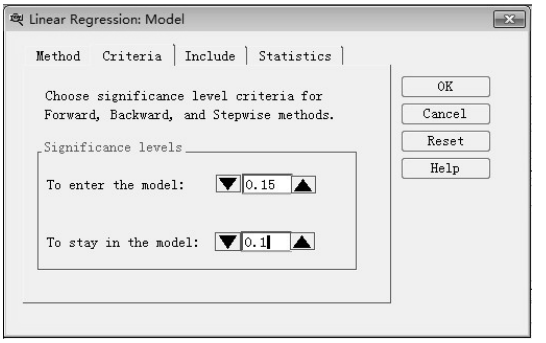


图 7-14　选择标准

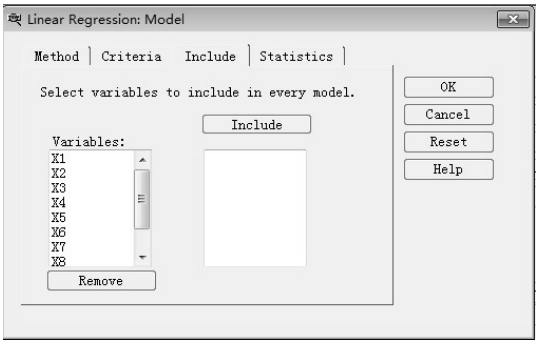


图 7-15　设置必选自变量

单击 Statistics（统计量）标签，显示如图 7-16 所示选项卡，注意到本选项卡中的选项仅在使用 Mallow's Cp 统计量选择法、R^2 选择法和校正 R^2 选择法进行模型选择时使用。Model fit

statistics（模型拟合统计量）选项框内包括 9 个准则：Akaike's information criterion（AIC 准则）、Bayesian information criterion（BIC 准则）、Schwarz's bayesian criterion（Schwarz's bayesian 准则）、Amemiya's prediction criterion（Amemiya's 预测准则）、Mallows' Cp statistic（Mallows' Cp 统计量）、Adjusted R-square（校正拟合优度）、Error sum of squares（误差平方和）、Error mean square（均方误）、Root mean square error（均方根）。因本例采取的为逐步回归模型选择法，在此不选择，单击 Cancel 按钮返回如图 7-12 所示对话框。

单击 Statistics（统计量）按钮，弹出如图 7-17 所示对话框，在 Parameter estimates（参数估计）选项框内可以定义参数估计的类型：Std.regression coefficients（标准回归系数）、Confidence limits for estimates（参数的置信区间）、Type 1 sum of squares（I 型平方和）、Type 2 sum of squares（II 型平方和）、Correlation matrix of estimates（估计值的相关矩阵）、Covariance matrix of estimates（估计值的协方差矩阵）。在 Correlations（相关分析）选项框内可选择 Partial correlations（偏相关分析）和 Semi-partial correlations（半-偏相关分析）。勾选 Std.regression coefficients（标准回归系数）、Type 1 sum of squares（I 型平方和）和 Type 2 sum of squares（II 型平方和）。

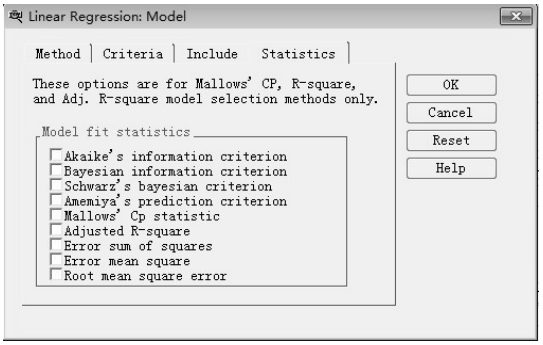


图 7-16　选择输出统计量

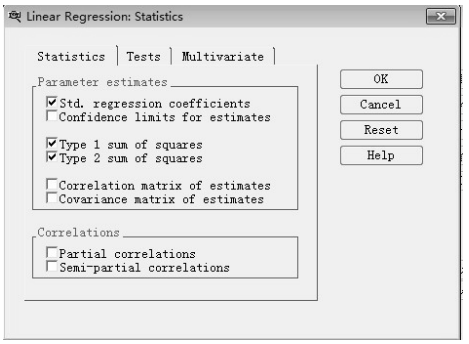


图 7-17　参数估计

单击 Tests（检验）标签，则显示如图 7-18 所示选项卡。勾选 Collinearity（共线性诊断）选项框内的 Collinearity analysis（共线性分析）、Tolerance values for estimates（估计值的容忍度）、Variance inflation factors（方差膨胀因子）三个项目。勾选 Autocorrelation（自相关分析）选项框内的 Durbin-Watson statistic（DW 统计量）选项。若需要进行 Heteroscedasticity（异方差分析）检验，可勾选 Heteroscedasticity（异方差分析）和 Asymptotic covariance matrix（渐进协方差矩阵）选项。

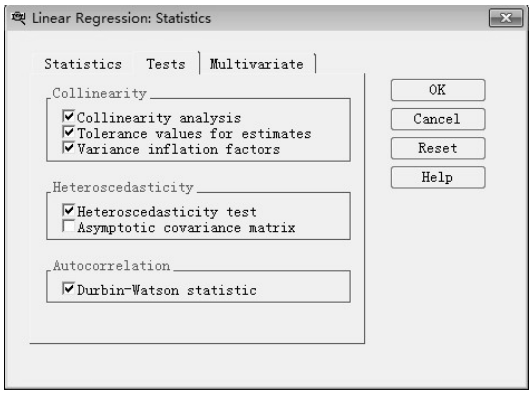


图 7-18　共线性检验

单击 Plots（绘图）按钮。在选项卡 Predicted（预测值）、Residual（残差值）和 Influence（影响统计量）内分别可设置绘制预测图、残差图和影响图。本例不设置，单击 Cancel 按钮返回如图 7-12 所示对话框。

单击图 7-12 所示对话框上的 OK 按钮，将在 Analyst 结果输出窗口显示结果目录树。单击相应选项即可打开结果，菜单操作得到的建模结论和编程得到的结果一致。

7.3　非线性回归

7.3.1　基本原理

现实世界中严格的线性模型并不多见，它们或多或少都带有某种程度的近似；在很多情况下，非线性模型可能更加符合实际。对变量间非线性相关问题的曲线拟合，处理方式主要有：

- 决定非线性模型的函数类型，对于其中可线性化问题则通过变量变换将其线性化，从而归结为前面的多元线性回归问题来解决。表 7-24 列出了部分常见的将非线性函数转换成线性函数的数据转换方法。
- 若实际问题的曲线类型不易确定时，由于任意曲线皆可由多项式来逼近，故常可用多项式回归来拟合曲线。
- 若变量间非线性关系式已知（多数未知），且难以用变量变换法将其线性化，则进行数值迭代的非线性回归分析。

表 7-24　典型的函数及线性化方法

函 数 名 称	函数表达式	线性化方法
双曲线函数	$\frac{1}{y}=a+\frac{b}{x}$	$v=\frac{1}{y}$　　$u=\frac{1}{x}$
幂函数	$y=ax^b$	$v=\ln y$　　$u=\ln x$
指数函数	$y=ae^{bx}$	$v=\ln y$　　$u=x$
	$y=ae^{b/x}$	$v=\ln y$　　$u=\frac{1}{x}$
对数函数	$y=a+b\ln x$	$v=y$　　$u=\ln x$
S 型函数	$y=\frac{1}{a+be^{-x}}$	$v=\frac{1}{y}$　　$u=e^{-x}$

7.3.2　SAS 过程——NLIN 过程

NLIN 过程的功能主要是计算非线性模型参数的最小二乘估计 LS 及加权最小二乘估计。NLIN 过程的一般使用格式如下：

```
PROC  NLIN    DATA=数据集   </选项列表>;
PARAMETERS  参数名=数值;
MODEL         因变量＝表达式  </选项列表>;
```

```
BOUNDS      表达式;
DER.参数名｛. 参数名｝= 表达式;
ID          变量列表;
OUTPUT      OUT=数据集 </选项列表>;
BY          变量列表;
RUN;
```

调用 NLIN 过程时必须定义 PARAMETERS 语句和 MODEL 语句，其他语句供选。

PROC NLIN 语句后的主要控制选项如表 7-25 所示。

表 7-25　PROC NLIN 语句后的主要控制选项

选　项	意　义
OUTEST＝数据集名	指定存放参数估计的每步迭代结果的数据集名
METHOD＝GAUSS | MARQUARDT | NEWTON| GRADIENT| DUD |	设定参数估计的迭代方法。默认时为 GAUSS
EFORMAT	所有数值以科学记数法输出
NOPOINT	抑制打印输出
NOINPOINT	抑制迭代结果的输出

NLIN 过程中主要使用的语句含义如下：

PARAMETERS（PARMS）语句——用于参数赋初值，项目间用空格分隔。例如，语句：PARMS B0=0 B1=1 TO 10 B2=1 TO 10 BY 2 B3=1,10,100;。

MODEL 语句——此语句定义的表达式可为任意有效 SAS 表达式，包括参数名字、输入数、据集中的变量名或在 NLIN 过程中用程序设计语句创建的新变量。例如：

```
MODEL Y=B0*(1－EXP(-B1*X));
```

BOUNDS 语句——用于设定参数取值范围（主要为不等式），约束间用逗号分隔。例如：

```
BOUNDS A<=20，B>30，1<=C<=10;
```

DER.语句——DER.语句用于计算模型关于各个参数的偏导数，相应的格式为：

```
一阶偏导数　DER．参数名＝表达式;
二阶偏导数　DER．参数名．参数名＝表达式;
```

如对于 MODEL Y=B0*(1－EXP(-B1*X))；DER.语句的书写格式为：

```
DER.B0=1－EXP(-B1*X);
DER.B1=B0*X*EXP(-B1*X);
```

对于多数算法，都必须对每个被估计的参数给出一阶偏导数表达式。对于 NEWTON 法，必须给出一、二阶偏导数表达式。

OUTPUT 语句——用于把一些计算结果输出到指定的数据集中。

7.3.3　SAS 实例——拟合某微生物生长曲线

例 7-4　已知某种微生物在一定温度下随时间变化的平均增长倍数数据如表 7-26 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap7\microbes”）。试拟合这种微生物的生长曲线。

表 7-26　某微生物生长数据

时　　间	1	2	3	4	5	6	7	8	9
增长倍数	1.3	1.5	2.6	3.6	6.8	8.4	8.5	9.1	9.5

解析：在分析此类问题时，一般首先绘制散点图或折线图，大致判断出曲线的类型，然后再分别拟合可能的曲线，最终根据判别系数选择最合适的回归模型拟合数据。

编程法：

（本例完整程序在光盘中的存储路径为“proc\chap7\microbes”。）

首先编写如下程序绘制这种微生物平均增长倍数随着时间点移动的折线图：

```
proc gplot data=chap7.microbes;
plot growth*time;
symbol    v=dot i=l c=black;                    /*指定用黑色星形表示数据、数据点用折线连接*/
run;
```

选择 Run|Submit 命令提交程序，则得到的折线图如图 7-19 所示。

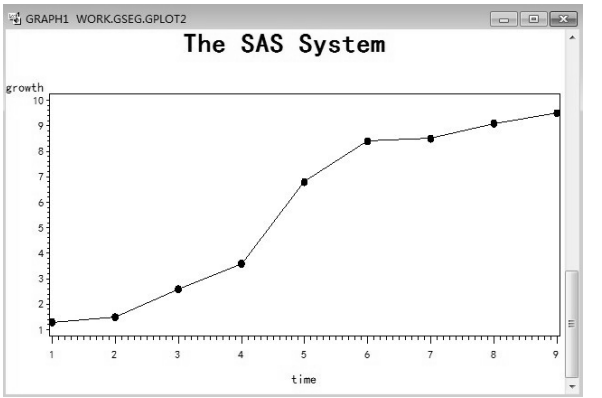


图 7-19　时间点与生长曲线的折线图

观察图 7-19 可初步得到该曲线呈现“S”形，估计其为三次曲线。编写如下程序分别拟合直线、抛物线和三次曲线，选择最优函数拟合，拟合前两种回归模型作为比较：

```
/*新建临时数据集 temp，导入数据集 chap7.microbes，并定义新变量 time2 和 time3*/
data temp;
set chap7.microbes;
time2=time*time;
time3=time*time*time;
run;

/*定义三个回归模型，并进行残差分析和误差项相关检验*/
proc reg data=temp;
model growth=time/dw r ;
model growth=time time2/r dw;
model growth=time time2 time3/r dw;
run;
ods graphics off;
```

选择 Run|Submit 命令提交程序，以下分别解析直线回归、抛物线和三次曲线的拟合结果。

1．模型 $y=b_0+b_1x$ 拟合结果

表 7-27 为模型的方差分析表，对模型的 *F* 检验 *P* 值小于 0.0001，则模型是显著的。表 7-28 为模型拟合的评价信息，其中拟合优度为 0.9306，校正的拟合优度为 0.9207。表 7-29 为模型的回归系数估计，对截距和变量 cost 的原假设“此系数为零”的 *t* 检验得到的 *P* 值分别为 0.6642 和小于 0.0001，则不能拒绝“截距项为零”的原假设，但是变量 time 是显著存在模型中的。

表 7-27　模型一的方差检验结果

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	86.88067	86.88067	93.86	<0.0001
Error	7	6.47933	0.92562		
Corrected Total	8	93.36000			

表 7-28　模型一拟合参数

Root MSE	0.96209	R-Square	0.9306
Dependent Mean	5.70000	Adj R-Sq	0.9207
Coeff Var	16.87879		

表 7-29　模型一参数估计

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	−0.31667	0.69894	−0.45	0.6642
time	time	1	1.20333	0.12421	9.69	<0.0001

表 7-30 为对误差项的相关性检验，DW 统计量的值为 1.147，误差项可能存在正相关。

表 7-30　误差项 DW 检验

Durbin-Watson D	1.147
Number of Observations	9
1st Order Autocorrelation	0.334

图 7-20 为残差诊断结果，观察左上角第一、第二个残差图，明显看出残差分布非随机。观察左排第二个关于残差的 QQ 图和第三个残差分布的直方图，初步得到残差分布也不满足正态的结论。

综上所述，虽然由表 7-27 得到模型为显著的，但是由 DW 检验结果可知本例数据不满足误差项独立的假设条件，且由分析残差图可知不满足误差项随机、为正态分布的假设条件。因此本例数据不适合拟合直线回归模型。

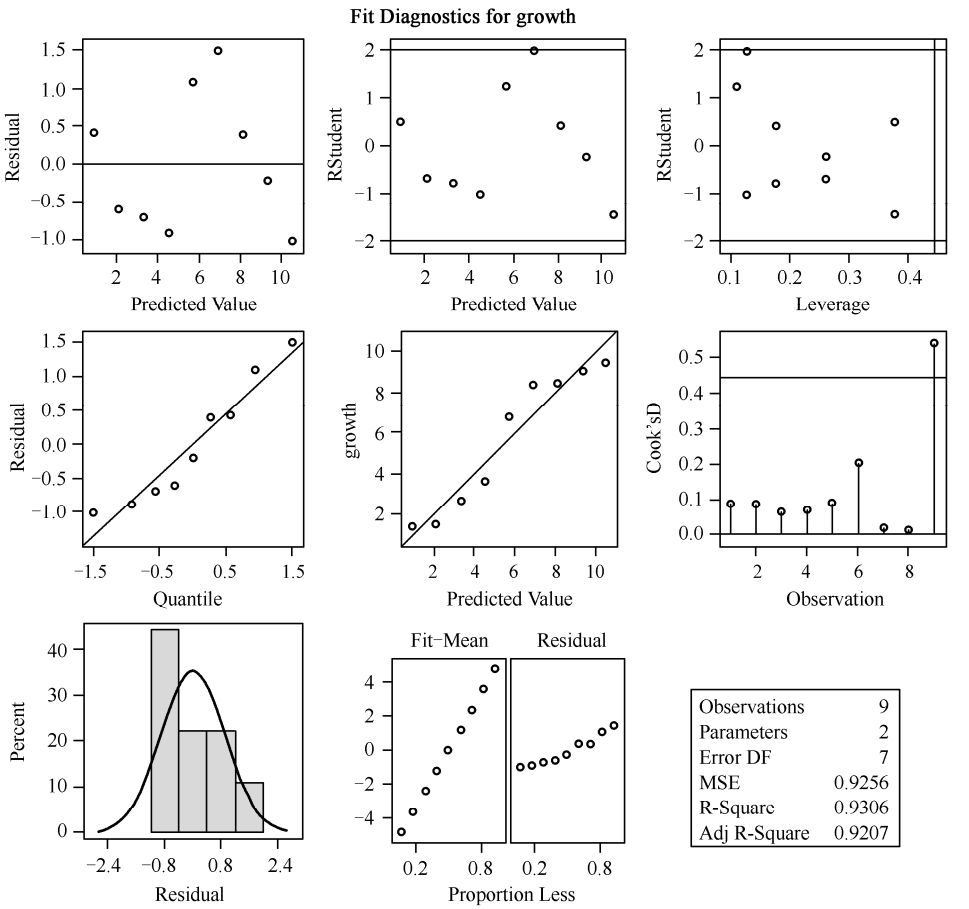


图 7-20　模型一残差诊断图

2．模型 $y=b_0+b_1x+b_2x^2$ 拟合结果

对模型的 *F* 检验 *P* 值小于 0.0001（表 7-31），则模型显著成立。表 7-32 为模型的拟合评价信息表。模型的拟合优度达到了 0.9421，校正后的拟合优度也达到了 0.9215，说明模型拟合度较高。表 7-33 为模型的回归系数估计信息，并对估计值进行原假设“改系数值为零”的 *t* 检验，截距项对应 *t* 检验的 *P* 值为 0.3093，自变量 cost 对应的 *t* 检验 *P* 值分别为 0.0195 和 0.3402，则在 0.05 的显著性水平之下，不能拒绝“该参数为零”的原假设；而变量 cost2 对应的 *t* 检验 *P* 值为 0.3402，则在 0.05 的显著性水平之下，不能拒绝 cost2 前对应的“系数为零”的原假设。

表 7-31　模型二方差检验结果

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	87.86365	43.93183	47.96	0.0002
Error	6	5.49635	0.91606		
Corrected Total	8	93.36000			

表 7-32　模型二拟合参数

Root MSE	0.95711	R-Square	0.9411
Dependent Mean	5.70000	Adj R-Sq	0.9215
Coeff Var	16.79139		

表 7-33　模型二参数估计

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	−1.35238	1.21784	−1.11	0.3093
time	time	1	1.76827	0.55919	3.16	0.0195
time2		1	−0.05649	0.05454	−1.04	0.3402

表 7-34 为模型的误差项相关性检验，DW 统计量的值为 1.338，则满足误差项可能存在正相关。

表 7-34　误差项 DW 检验

Durbin-Watson D	1.338
Number of Observations	9
1st Order Autocorrelation	0.229

图 7-21 为残差诊断结果，观察左上角第一、第二个残差图，明显看出残差分布非随机。观察左排第二个关于残差的 QQ 图和第三个残差分布的直方图，初步得到残差分布也不满足正态的结论。

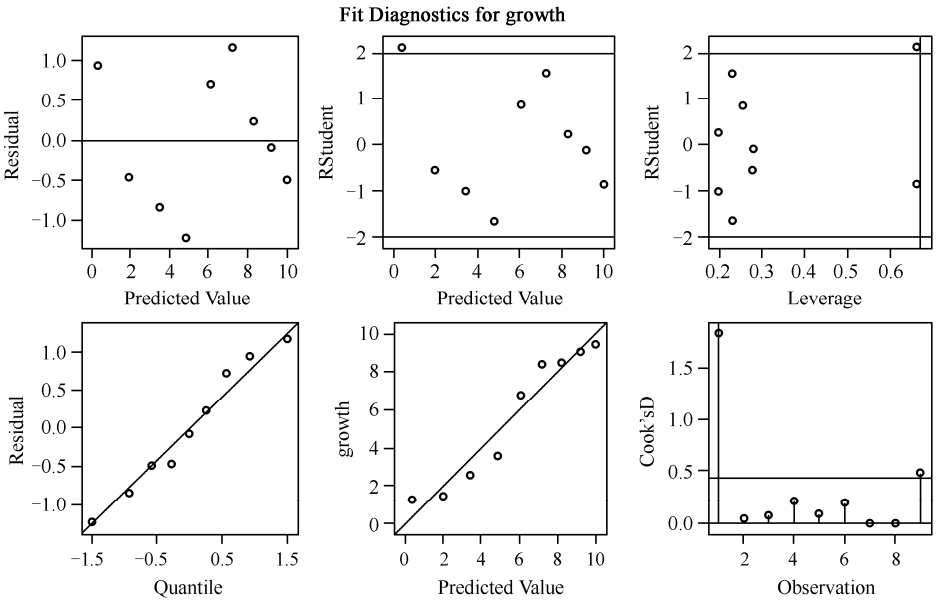


图 7-21　模型二残差诊断图

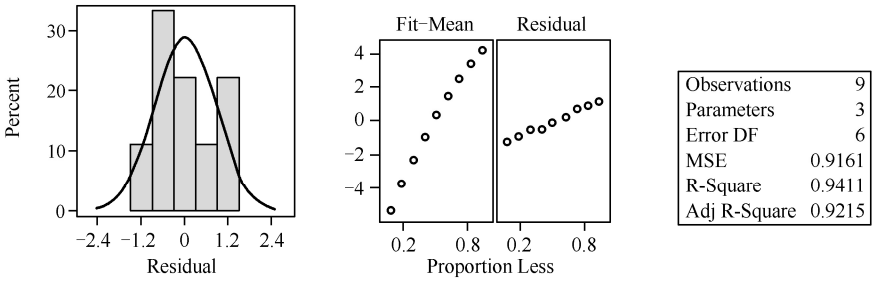


图 7-21　模型二残差诊断图（续）

综上所述，虽然由表 7-31 得到模型为显著的，但是模型中无论截距项还是各变量前的参数都不显著非零，而且由 DW 检验结果可知本例数据不满足误差项独立的假设条件，而由分析残差图可知不满足误差项随机、为正态分布的假设条件。因此本例数据不适合拟合抛物线模型。

3．模型 $y=b_0+b_1x+b_2x^2+b_3x^3$ 拟合结果

对模型的 *F* 检验 *P* 值等于 0.0001（表 7-35），模型是显著成立的。其中拟合优度 R-Square 和校正拟合优度 Adj R-Sq 分别为 0.9779 和 0.9646（表 7-36），则模型拟合度较高。表 7-37 为回归系数参数估计与检验，截距项对应的 *t* 检验 *P* 值为 0.2324，变量 time 对应的 *t* 检验 *P* 值为 0.2945，即不能拒绝“截距项（变量 time 前的系数）为零”的原假设，考虑拟合删除这两项的回归模型。

表 7-35　模型三方差检验结果

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	91.29688	30.43229	73.75	0.0001
Error	5	2.06312	0.41262		
Corrected Total	8	93.36000			

表 7-36　模型三拟合参数

Root MSE	0.64236	R-Square	0.9779
Dependent Mean	5.70000	Adj R-Sq	0.9646
Coeff Var	11.26945		

表 7-37　模型三参数估计

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1.88651	1.38883	1.36	0.2324
time	time	1	−1.33321	1.13883	−1.17	0.2945
time2		1	0.67962	0.25781	2.64	0.0462
time3		1	−0.04907	0.01701	−2.88	0.0344

由于截距项和变量 time 前的系数都不显著非零，因此编写如下程序拟合去掉这两项的模型：

```
/*修改三次曲线回归模型，并进行残差分析和误差项相关检验*/
proc reg data=temp;
model growth=time2 time3/r dw noint;
run;
```

选择 Run|Submit 命令提交程序，以下分析主要输出结果。

对模型的 *F* 检验 *P* 值小于 0.0001（表 7-38），模型是显著成立的。其中拟合优度 R-Square 和校正拟合优度 Adj R-Sq 分别为 0.9926 和 0.9905（表 7-39），则模型拟合度高，预示着自变量对因变量的预测作用强。表 7-40 为回归系数参数估计与检验，变量 time2 和 time3 对应的系数均显著非零（对应的 *t* 检验 *P* 值小于 0.001）。

表 7-38　模型的方差检验结果

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	382.90501	191.45250	467.77	<0.0001
Error	7	2.86499	0.40928		
Uncorrected Total	9	385.77000			

表 7-39　模型拟合参数

Root MSE	0.63975	R-Square	0.9926
Dependent Mean	5.70000	Adj R-Sq	0.9905
Coeff Var	11.22375		

表 7-40　参数估计值

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
time2		1	0.41409	0.03151	13.14	<0.0001
time3		1	−0.03327	0.00395	−8.43	<0.0001

表 7-41 所示的对误差项的相关性检验结果表明，因 DW 统计量为 1.722，接近于 2，则满足误差项独立的假设条件。

表 7-41　误差项 DW 检验

Durbin-Watson D	1.722
Number of Observations	9
1st Order Autocorrelation	−0.017

图 7-22 为残差诊断结果，观察左上角第一、第二个残差图，看出残差分布是随机的。观察左排第二个关于残差的 QQ 图和第三个残差分布的直方图，得到残差分布近似满足正态分布。

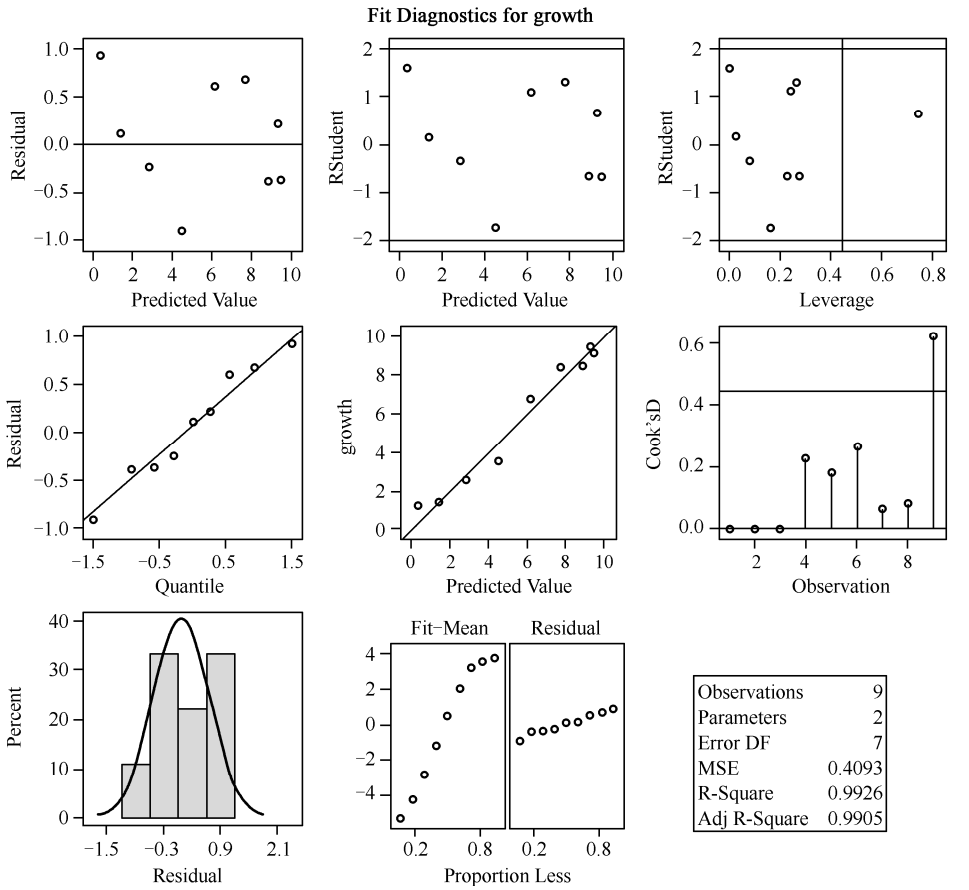


图 7-22　残差诊断图

综上所述，本例中拟合调整后的第三个模型无论是模型的显著性、对回归参数估计值的 *t* 检验，还是误差项的自相关检验及模型的拟合优度都能得到比较让人满意的结果。因此最终确定使用调整后的模型三，拟合的模型为：

$$\text{Growth} = 0.41409\text{time}^2 - 0.03327\text{time}^3$$

菜单法：

步骤一：选择 Solutions|Analysis|Analyst 命令，进入 Analyst 分析主界面。

步骤二：选择 File|Open By SAS Name|chap7|Microbes|OK 命令，打开数据集 chap7.Microbes。

步骤三：选择 Statistics|Regression|Simple 命令，弹出如图 7-23 所示对话框，单击变量 growth，再单击选项 Dependent（因变量），则将变量 growth 选为回归模型的因变量。用同样的方式将变量 time 选为 Explanatory（自变量）。单击选择 Model（模型）选项框中的 Cubic 选项，则设置建立三次抛物线回归模型。

单击 Statistics（统计量）按钮，弹出如图 7-24 所示对话框，勾选 Parameter estimates（参数估计）选项框中的 Std.regression coefficients（标准回归系数）、Confidence limits for estimates（参数的置信区间）选项。单击 OK 按钮保存设置并返回如图 7-23 所示对话框。

单击 Predictions（预测）按钮，弹出如图 7-25 所示对话框，勾选 Predict original sample（对原始样本进行预测）和 List predictions（以列表形式输出预测值）选项，单击 OK 按钮保存设置并返回如图 7-23 所示对话框。

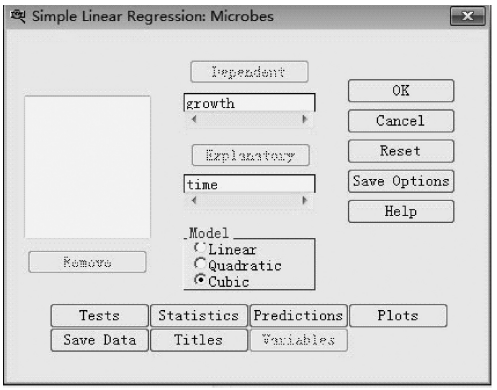


图 7-23　简单回归分析主对话框

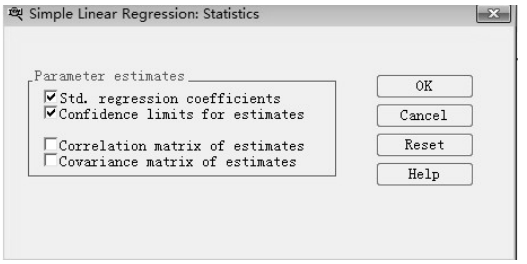


图 7-24　参数估计设置

单击图 7-23 所示界面上的 OK 按钮，将在 Analyst 结果输出窗口显示结果目录树。单击相应选项即可打开结果。此输出结果和方法一中建立的模型二中结果一致。读者可参照以上结果自行分析。

注意：在图 7-23 所示的主界面中单击 Plots 按钮可在弹出的对话框中设置绘制图形；单击 Save Data 按钮可在弹出的对话框中设置保存分析指标；单击 Titles 按钮可在弹出的对话框中设置分析结果标题。

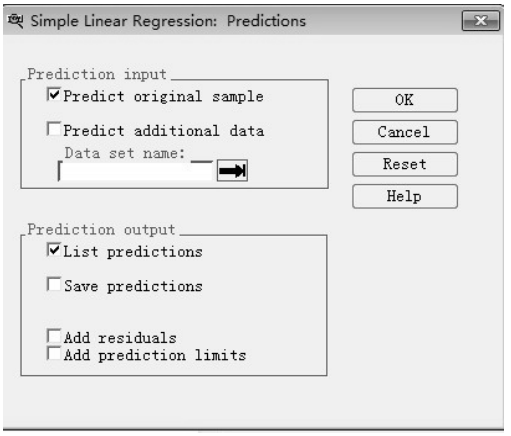


图 7-25　设置预测值

7.3.4　SAS 实例——非线性回归函数的参数估计

例 7-5　已知关于基底和反应速度的酶动力学理论公式为：$f(x,\theta_1,\theta_2)=\dfrac{\theta_1 x_i}{\theta_2+x_i}\ (i=1,2,\cdots,n)$，其中 x_i 代表基低的量，$f(x,\theta_1,\theta_2)$ 代表反应速度。请根据表 7-42 所示试验数据（相应的 SAS 数据集在光盘中的存储路径为“data\chap7\enzyme”）估计参数 θ_1 和 θ_2，并探索该数据用理论公式拟合是否合理？

表 7-42　酶催化效率试验数据

基底量（concentration）	反应速度（velocity）	基底量（concentration）	反应速度（velocity）
0.26	124.7	1.14	147.6
0.3	126.9	1.28	149.8
0.48	135.9	1.38	149.4
0.5	137.6	1.8	153.9
0.54	139.6	2.3	152.5
0.68	141.1	2.44	154.5
0.82	142.8	2.48	154.7

解析：酶动力学公式为典型的非线性模型，以下采用 NLIN 过程进行非线性模型中的参数估计。编写如下程序（其在光盘中的存储路径为“proc\chap7\enzyme”）：

```
proc nlin data=chap7.enzyme;                    /*调用 nlin 过程*/
parms theta1=1 theta2=1;
/*定义两个参数 theta1、theta2,并给出它们迭代的初始值*/
model velocity=theta1*concentration/(theta2+concentration);
run;
```

选择 Run|Submit 命令提交程序，以下分析主要的输出结果。表 7-43 为迭代的详细信息，在本次非线性回归中进行了 21 次迭代，最终让平方和稳定于 244637。

表 7-43　迭代信息

Iterative Phase			
Iter	theta1	theta2	Sum of Squares
0	1.0000	1.0000	288160
1	1.1831	−1.1116	273522
2	1.5306	−1.1046	272096
3	3.0623	−1.0763	267520
4	6.4367	−1.0556	256749
5	6.8208	−1.0915	247819
6	13.7225	−1.0433	247290
7	9.4027	−1.0891	246646
8	14.4583	−1.0530	246170
9	10.8041	−1.0784	245085
10	12.8774	−1.0605	245059
11	10.9059	−1.0750	244755
12	11.5645	−1.0694	244648
13	11.2056	−1.0712	244642
14	11.3770	−1.0698	244638
15	11.3032	−1.0703	244637

续表

Iterative Phase			
Iter	theta1	theta2	Sum of Squares
16	11.3233	−1.0701	244637
17	11.2987	−1.0702	244637
18	11.3067	−1.0702	244637
19	11.3045	−1.0702	244637
20	11.3047	−1.0702	244637
21	11.3042	−1.0702	244637

表 7-44 为模型的方差分析表，模型进行 F 检验得到的近似 P 值为 0.3594，说明模型在显著性水平 0.05 下不是显著成立的。

表 7-44　模型方差检验

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	2	45498.5	22749.3	1.12	0.3594
Error	12	244637	20386.4		
Uncorrected Total	14	290135			

表 7-45 为系统应用极大似然法算出的参数 θ_1 和 θ_2 的估计值、标准误和 95%置信区间，θ_1（theta1）=11.3042、θ_2（theta2）=−1.0702，则得到函数形式：

$$\text{velocity}=\frac{11.3042\text{concentration}}{-1.0702+\text{concentration}}$$

表 7-45　参数估计

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
theta1	11.3042	17.4161	−26.6422	49.2506
theta2	−1.0702	0.1229	−1.3379	−0.8025

本例虽然根据试验数据基于酶催化理论公式得到了相应的函数式，但是根据模型的方差检验结果，可知该数据并不适合用理论催化公式进行拟合，需根据实际进行调整，感兴趣的读者可自己尝试变换函数形式。

7.4　LOGISTIC 回归

7.4.1　基本原理

在某些回归问题中因变量是分类变量的，它的取值可能是态度（同意、反对），或者疾病治疗后治愈或复发的可能，或者是有序变量（如 0=低骨密度、1=正常、2=高骨密度）。当因变量

为二分类变量情形，令 y_i 为 n_i 次试验中测得成功的次数，再假定 y_i 是一个在 n_i 次试验中的二项随机变量，其中任何一次试验成功的概率为 θ_i，在 LOGISTIC 回归中，以 θ_i 作为因变量建立回归模型，对因变量进行如下所示的 logit 变换：

$$\text{logit}(\theta_i)=\ln\left(\frac{\theta_i}{1-\theta_i}\right)$$

由此 LOGISTIC 回归可写作以下两种形式：

$$\text{logit}(\theta_i)=\beta_0+\beta_1 x_i$$

$$E(\tfrac{y_i}{n})=\theta_i=\frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}$$

二分类变量的 LOGISTIC 回归模型包括三要素：

（1）因变量为在已知试验次数的试验中的成功次数构成的相互独立的二项计数组成的；

（2）成功的概率只以线性的形式依赖于自变量；

（3）存在一个函数（在此为 logit 函数）将自变量的线性形式和二项计数的期望值相联系。

在 SAS 系统中，LOGISTIC 过程可用如下方法建立回归模型：全回归模型、前向选择法、后向淘汰法、逐步回归法、最佳子集法。

7.4.2　SAS 过程——LOGSTIC 过程

LOGISTIC 过程适合处理因变量为二分或二分以上的类别数据，当模型中的自变量数目过多时，LOGISTIC 程序还可提供逐步筛选的方法来挑选最佳模型。它的一般使用格式如下：

```
PROC LOGISTIC   DATA=SAS 数据集 <选项列表>;
MODEL 因变量=自变量</选项列表>;
OUTPUT =输出文件名 关键字=变量名称/ALPHA=概率值;
WEIGHT 变量名;
BY 变量名;
RUN;
```

调用 LOGISTIC 过程时必须定义 PROC LOGISTIC 语句和 MODEL 语句。PROC LOGISTIC 语句后主要控制选项如表 7-46 所示。

表 7-46　PROC LOGISTIC 语句后主要控制选项

选　项	意　义
ORDER=DATA\INTERNAL\ FORMATTED	指定因变量取值组别的次序，若 ORDER=DATA，则按输入文件内各组出现的次序排列；若 ORDER=INTERNAL，则按因变量值的大小或字母排序；若 ORDER=FORMATTED，则组别次序由外在格式决定。系统默认 ORDER=INTERNAL
NOSIMPLE	不输出所有自变量的描述性统计量
NOPRINT	不输出任何回归分析结果

MODEL 语句——定义 LOGISTIC 回归的模型，此语句中的因变量为二分的名义变量或次序变量；定义的自变量必须是连续性变量。若不定义自变量，则系统认为回归模型中只含截距参数。MODEL 语句后主要的控制选项如表 7-47 所示。

表 7-47　MODEL 语句后主要控制选项

选　项	意　义
NOINT	在建立回归模型时不考虑截距参数
SELECTION=NONE\FORWARD BACKWARD\STEPWISE\SCORE	定义选择最佳回归模型的方法。其中 NONE 表示将所有自变量纳入模型；FORWARD 代表前向选择法；BACKWARD 代表反向淘汰法；STEPWISE 代表逐步选择法；SCORE 代表根据最大可能分数来界定最佳模型
ALPHA=	定义计算置信区间的显著性水平
CLPARM=	计算估计参数的置信区间
INFLUENCE	输出异常点的诊断
RSQUARE	输出判别系数 R^2

WEIGHT 语句——定义观测在分析中的加权值。

BY 语句——定义分析的分层变量，前提是数据集预先按分层变量排序。

7.4.3　SAS 实例——结石病危险因素研究

例 7-6　某研究小组对某村进行健康调查，随机抽取了 300 个大于 15 周岁的居民，以问卷调查的方式得到表 7-48 所示的调查数据（包含完整数据的 SAS 数据集在光盘中的存储路径为“data\chap7\disease”）。

disease：是否患慢性病（1=是；0=否）。

sex：性别（1=男；2=女）。

age_type：年龄阶段（1=15~25；2=25~35；3=35~45；4=45~55；5=55~65；6=65）。

marriage：婚姻状况（1=“未婚”；2=“初婚”；3=“再婚”；4=“离婚”；5“丧偶”）。

education：文化程度（1=“不识字或识字少”；2=“小学”；3=“初中”；4=“高中、中专”；5=“大专、本科”；6=“其他”）。

smoking：吸烟（1=“不吸烟”；2=“以前吸现在不吸”；3=“一直吸”）。

试分析这些因素是否为居民患慢性病的影响因素？

表 7-48　居民健康调查数据

ID	sex	marriage	education	smoking	age_type	disease
1	1	2	3	3	5	1
2	2	2	4	1	5	1
3	2	2	3	1	2	0
4	1	2	3	1	4	0
5	2	2	3	1	3	0
6	1	2	3	3	3	0
7	2	1	2	3	6	0
8	1	1	2	3	5	0
9	2	5	1	1	6	0

解析：本例为典型的应用 logstic 回归模型探索影响因素，包括影响居民患慢性病的危险因素和保护因素等。

编程法：

编写如下程序（其在光盘中的存储路径为“proc\chap7\disease”）：

```
ods graphics on;
proc logistic data=chap7.disease;       /*调用 logistic 过程*/
class sex marriage education smoking age_type;
model disease(event='1')=sex marriage education smoking age_type
                        / selection=stepwise
                          slentry=0.15    /*定义 logistic 回归模型，指定用逐步回归方法选择模型，
                          slstay=0.15     变量被选进和剔除模型的显著性水平分别为 0.15 和 0.15*/
                          ;
run;
```

选择 Run|Submit 命令提交程序，表 7-49 列出了建模的基本信息，观察可知本例采用了二分类变量模型。由于本例的自变量都是分类变量，表 7-50 给出了这些自变量的重新编码信息。

表 7-49　模型信息

Model Information		
Data Set	CHAP7.DISEASE	
Response Variable	disease	disease
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

表 7-50　分类变量编码信息

Class Level Information						
Class	Value	Design Variables				
sex	1	1				
	2	−1				
marriage	1	1	0	0		
	2	0	1	0		
	4	0	0	1		
	5	−1	−1	−1		
education	1	1	0	0	0	
	2	0	1	0	0	
	3	0	0	1	0	
	4	0	0	0	1	
	5	−1	−1	−1	−1	
smoking	1	1	0			

续表

Class Level Information						
Class	Value	Design Variables				
	2	0	1			
	3	−1	−1			
age_type	1	1	0	0	0	0
	2	0	1	0	0	0
	3	0	0	1	0	0
	4	0	0	0	1	0
	5	0	0	0	0	1
	6	−1	−1	−1	−1	−1

由于本例设定的变量进入模型须达到显著性水平为 0.15，因此只有 age 一个变量被选入模型，以下分析建模结果。观察表 7-51 可知本模型为收敛的。表 7-52 列出了加入协变量（age）前后模型的 AIC、SC 和−2logL 的值，观察可知引入协变量以后模型的三个拟合指标值都减少，说明引入协变量的模型优于仅包含截距项的模型。表 7-53 为模型的系数整体检验结果，观察可知模型系数检验的多变量检验 *P* 值都小于 0.05，即拒绝“自变量前系数为零”的原假设。

表 7-51　模型收敛状态

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

表 7-52　模型拟合信息

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	315.594	287.379
SC	319.298	309.602
−2 log L	313.594	275.379

表 7-53　模型整体系数检验

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	38.2150	5	<0.0001
Score	32.6490	5	<0.0001
Wald	22.9810	5	0.0003

表 7-54 为系数的极大似然估计结果，包括参数估计及对应的卡方检验。表 7-55 为相对危险率的估计值。观察可知道，年龄阶段取值为 2（即 25～35 岁）的人群患慢性病危险率是年龄阶

段取值为 1（即 15～25 岁）的 0.253 倍，即患病率相对较低，而年龄阶段取值为 6（即 65 岁以上）的居民患慢性病相对危险率是年龄阶段取值为 1（15～25 岁）的居民的 7.355 倍。即 35 岁以上居民，随着年龄增长，居民患慢性病的风险也随着增加。

表 7-54　参数估计

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	−1.6420	0.2217	54.8519	<0.0001
age_type	2	1	−2.2700	0.8539	7.0671	0.0079
age_type	3	1	0.0839	0.3412	0.0604	0.8059
age_type	4	1	1.0359	0.3262	10.0854	0.0015
age_type	5	1	0.9488	0.3551	7.1410	0.0075
age_type	6	1	1.0984	0.3282	11.2023	0.0008

表 7-55　相对危险率估计

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age_type 2 vs 1	0.253	0.025	2.532
age_type 3 vs 1	2.667	0.705	10.084
age_type 4 vs 1	6.909	1.868	25.560
age_type 5 vs 1	6.333	1.640	24.451
age_type 6 vs 1	7.355	1.982	27.288

菜单法：

步骤一：选择 Solutions|Analysis|Analyst 命令，进入 Analyst 分析主界面。

步骤二：选择 File|Open By SAS Name|chap7|disease|OK 命令，打开数据集 chap7.disease。

步骤三：选择 Statistics|Regression|Logistic 命令，弹出如图 7-26 所示对话框。单击变量 disease，再单击选项 Dependent（因变量），则将变量 disease 定义为回归模型的因变量。用同样的方式将变量 sex、marriage、education、smoking、age_type 选进 Class（分类变量）选项框。单击 Model Pr（模型概率）选项框后的按钮，在出现的下拉选项框中单击“1”，则设置分析因变量值为 1 的概率。

单击 Model（模型）按钮，弹出如 7-27 所示对话框，单击 Standard Models（模型类型）按钮，在弹出的下拉菜单中选择 Main effects only（只分析主效应）。

单击 Selection（模型选择）标签，则显示如图 7-28 所示选项卡，选择 Selection method（选择方法）选项框中的 Stepwise selection（逐步选择法）选项。

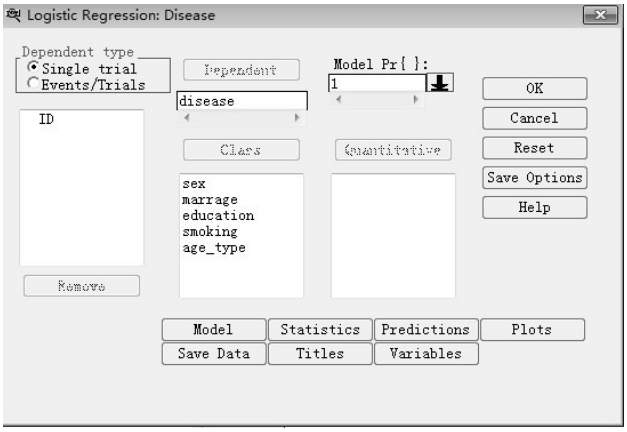


图 7-26　LOGISTIC 回归分析主对话框

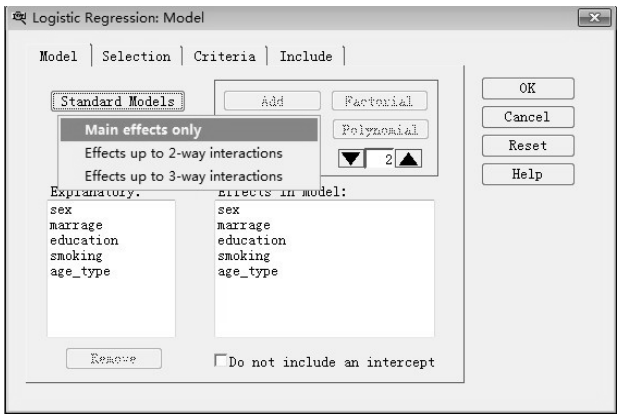


图 7-27　模型设置

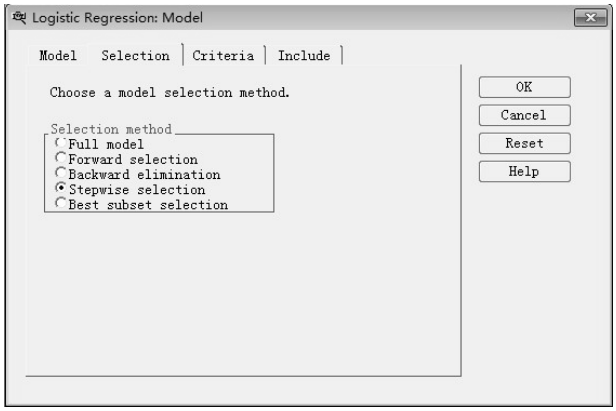


图 7-28　设置模型选择方法

单击 Criteria（标准）标签，则显示如图 7-29 所示选项卡，在 To enter the model（变量被选入模型的显著性水平）选项后填入 0.15，在 To stay in the model（变量被剔除模型的显著性水平）选项后填入 0.15。单击 OK 按钮保存设置并返回如图 7-26 所示对话框。单击 OK 按钮则输出和编程相类似的结果，请读者自行分析。

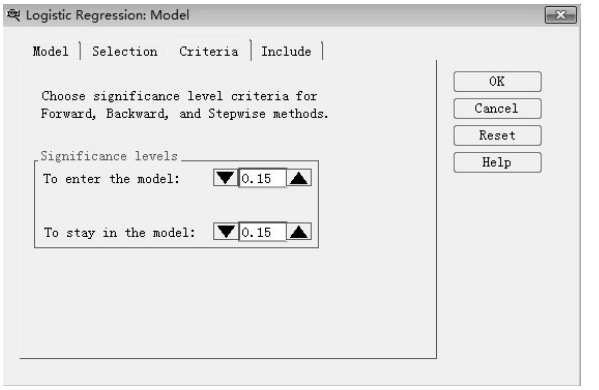


图 7-29　设置显著性水平

练习题

习题 7-1　某大学统计系中随机抽取了 20 名学生，对他们的数学分析（简称数分）和概率论成绩进行调查，得到结果如表 7-56 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap7\math”）。请绘制学生的数学分析与概率论成绩的散点图，并且计算它们之间的相关系数。

表 7-56　学生成绩记录

学生编号 (ID)	数分成绩 (math)	概率论成绩 (probability)	学生编号 (ID)	数分成绩 (math)	概率论成绩 (probability)
1	81	72	11	83	78
2	90	91	12	81	93
3	91	94	13	77	69
4	74	68	14	60	65
5	70	82	15	66	59
6	73	78	16	84	87
7	85	84	17	70	71
8	60	69	18	54	49
9	65	76	19	67	76
10	89	84	20	98	95

（本习题的解答程序在光盘中的存储路径为“proc\chap7\math”。）

习题 7-2　在一次体检中，记录了某班学生的年龄（Age）并测量得到他们的身高（Height）和体重（Weight），数据如表 7-57 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap7\class”）。

（1）请建立一元回归模型，用学生的身高预测体重；

（2）请建立多元回归模型，用学生的身高和年龄预测体重。

表 7-57　学生体检测量数据

姓名（Name）	身高（Height）	体重（Weight）	年龄（Age）	姓名（Name）	身高（Height）	体重（Weight）	年龄（Age）
Alfred	69	112.5	14	John	59	99.5	12
Alice	56.5	84	13	Joyce	51.3	50.5	11
Barbara	65.3	98	13	Judy	64.3	90	14
Carol	62.8	102.5	14	Louise	56.3	77	12
Henry	63.5	102.5	14	Mary	66.5	112	15
James	57.3	83	12	Philip	72	150	16
Jane	59.8	84.5	12	Robert	64.8	128	12
Janet	62.5	112.5	15	Ronald	67	133	15
Jeffrey	62.5	84	13	Thomas	57.5	85	11

（本习题的解答程序在光盘中的存储路径为“proc\chap7\class”。）

习题 7-3　为了检测出某汽车的最大耗油量，变换不同的车速同时记录下车子的耗油量，记录数据如表 7-58 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap7\mileage”）。试拟合车速和耗油量之间的函数。

表 7-58　不同车速和相应耗油量数据

车速（mph）	20	30	40	50	50	50	55	60	20
耗油量（mpg）	15.4	20.2	25.7	26.2	26.6	27.4		24.8	15.4

（本习题的解答程序在光盘中的存储路径为“proc\chap7\mileage”。）

习题 7-4　已知分段函数的表现形式为 $y=\begin{cases} a & x<x_0 \\ bx^2+cx+d & x\geqslant x_0 \end{cases}$，现在已知部分 x 和 y 的值如表 7-59 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap7\function”）。请估计未知参数 a、b、c 和 d。

表 7-59　分段函数部分取值

x	1	2	3	4	5	6	7	8
y	0.46	0.47	0.57	0.61	0.62	0.68	0.69	0.78
x	9	10	11	12	13	14	15	16
y	0.7	0.74	0.77	0.78	0.74	0.8	0.8	0.78

（本习题解答程序在光盘中的存储路径为“proc\chap7\function”。）

习题 7-5　某研究者为了探索摄入不同比率和体积的空气对血管的瞬时收缩的影响，在不同比率和体积空气的情形下进行了试验，数据如表 7-60 所示（相应的 SAS 数据集在光盘中的存储位置为“data\chap7\vaso”）。请拟合恰当的模型用空气比率和体积值预测血管瞬时收缩的发生。

表 7-60　不同比率和体积的空气对血管瞬时收缩的影响

比率（Rate）	体积（Volume）	结果（Response）	比率（Rate）	体积（Volume）	结果（Response）
3.7	0.825	constrict	0.4	2	no_constrict
3.5	1.09	constrict	0.95	1.36	no_constrict
1.25	2.5	constrict	1.35	1.35	no_constrict
0.75	1.5	constrict	1.5	1.36	no_constrict
0.8	3.2	constrict	1.6	1.78	constrict
0.7	3.5	constrict	0.6	1.5	no_constrict
0.6	0.75	no_constrict	1.8	1.5	constrict
1.1	1.7	no_constrict	0.95	1.9	no_constrict
0.9	0.75	no_constrict	1.9	0.95	constrict
0.9	0.45	no_constrict	1.6	0.4	no_constrict
0.8	0.57	no_constrict	2.7	0.75	constrict
0.55	2.75	no_constrict	2.35	0.03	no_constrict
0.6	3	no_constrict	1.1	1.83	no_constrict
1.4	2.33	constrict	1.1	2.2	constrict
0.75	3.75	constrict	1.2	2	constrict
2.3	1.64	constrict	0.8	3.33	constrict
3.2	1.6	constrict	0.95	1.9	no_constrict
0.85	1.415	constrict	0.75	1.9	no_constrict
1.7	1.06	no_constrict	1.3	1.625	constrict
1.8	1.8	constrict	0.4	2	no_constrict

注：结果列中取值 constrict 代表出现了血管瞬时收缩，no_constrict 代表没有出现血管瞬时收缩。

第 8 章　列联表分析

列联表分析主要用于离散型分类计数资料，检验数据是否服从特定的分布、离散或有序等级变量间是否存在相关关系、变量（评价）之间是否存在一致性等。这些分析主要包括建立在列联表基础上的拟合优度检验、独立性检验、一致性检验和计算属性变量关联度。

以下首先介绍 SAS 系统的 FREQ 过程，再结合实例以编程和菜单操作的方式具体介绍上述方法的实现。

8.1　SAS 过程——FREQ 过程

SAS 系统中的 FREQ 过程用于列联表分析，即分析包含有一个或多个类别变量的数据集文件。此过程主要功能如下：绘制次数分配表或列联表；计算 λ^2 统计量、Fisher's 精确性检验统计量（Fisher's Exact Test）等。

FREQ 语句的一般使用格式为：

```
PROC FREQ DATA=SAS 数据集 <选项列表>;
BY 变量列表;
WEIGHT 变量列表;
EXACT 选择统计量< /选项列表>;
TABLES 要求列表 < /选项列表>;
OUTPUT <OUT=数据集名> 选项列表;
RUN;
```

PROC FREQ 语句后的选项如表 8-1 所示。

表 8-1　PROC FREQ 语句后的选项列表

选　项	意　义
FORMCHAR（1，2，7）	定义列联表的边缘和中间分割线的形式，如 FORMCHAR（1，2，7）='|-+'代表列表的纵向用“|”表示，列表的横向用“-”表示，交叉处用“+”表示，此为系统默认形式
ORDER=	定义某一变量下各类别的输出次序，若定义 ORDER=FREQ，则次序先后按类别次数降序排列；若 ORDER=DATA，则类别次序即为它们在输入文件内出现的次序；若 ORDER=INTERNAL 时，类别次序由英文字母先后顺序决定；若 ORDER=FOUMATTED，则类别次序由外在格式决定。若此项省略，则系统默认按英文字母先后顺序输出，且缺失数据排在最前面
PAGE	每一页打印一张表格
NLEVELS	输出所有列表变量的水平数
NOPRINT	不打印出结果

FREQ 过程中主要使用的语句含义如下：

BY 语句——定义分层分析变量，要求分析数据集事先按 BY 语句指定的变量排序。

WEIGHT 语句——定义观测的加权变量（一般每个观测只代表一个数据点）。加权变量的值可包含小数，但必须为正。

EXACT 语句——使用 EXACT 语句需要精确定义指定统计量的检验或置信度，此语句适用于数据量较少的精确计算。其后可以定义的统计量主要如表 8-2 所示。

表 8-2　EXACT 语句可定义的统计量

统 计 量	意　义
AGREE	对二维列联表的尼曼检验，简单/加权的卡帕参数检验
BINOMAL	对一维表格的二项式比例检验
CHISQ	对一维表格的 λ^2 拟合检验；对二维列联表的 PEARSON λ^2 检验，似然比 λ^2 检验和 MANTEL-HAENSZEL λ^2 检验
FISHER	FISHER'S 精确检验
JT	Jonckheere-Terpsta 检验
KAPPA	简单的 KAPPA 参数检验
LRCHI	似然比 λ^2 检验
MCNEM	对 2*2 维列联表进行 MCNEMAR'S 检验
MEASURES	对 2*2 维列联表进行 PEARSON 系数和 SPEARMAN 系数检验，估计风险比的置信度
MHCHI	MANTEL-HAENSZEL λ^2 检验
OR	对 2*2 维列联表估计风险比的置信度
PCHI	PEARSON λ^2 检验
PCORR	检验 PEARSON 相关系数
RISKDIFF	对 2*2 维列联表的比率差的置信度，若估计第一列（第二列）的比率差的置信度可用 RISKDIFF1（RISKDIFF2）
SCORR	Spearman 相关系数检验
TREND	Cochran-Armitage 趋势检验
WTKAP	加权 KAPPA 系数检验

EXACT 语句斜杠（/）后可选的主要控制选项如表 8-3 所示。

表 8-3　EXACT 语句后主要的可选项

选　项	意　义
ALPHA=A	定义显著性水平
MAXTIME	定义 FREQ 过程每计算一个精确的 *P* 值的时间（单位为秒）
MC	要求用 MONTE-CARLO 方法估计每个 *P* 值，而非计算精确 *P* 值
N=n	定义 MONTE-CARLO 估计的最大次数，必须取整数，系统缺失值为 10000
POINT	对检验统计量进行点估计
SEED=NUMBER	定义随机 MONTE-CARLO 估计的原始种子，必须取整数

TABLES 语句——此语句主要用于设计频数分布表，即安排一元、二分类或多分类变量。

定义时，一般用星号连接各变量，常见定义形式如表 8-4 所示。

表 8-4　TABLES 语句一般定义形式

形　式	含　义
TABLES A B;	画次数分布表
TABLES A*B/选项;	画二维交叉表，变量 A 为行，变量 B 为列
TABLESA*B*C/选项;	画三维交叉表，变量 A 形成表的层，变量 B 为行，变量 C 为列

若需定义多个频数分布表，可用括号及两横线（--）简化语句撰写，表 8-5 列出了几种简化的定义形式。

表 8-5　TABLES 语句简化定义形式

简 化 形 式	等 价 形 式
TABLESA*(B C)	TABLES A*B A*C;
TABLES(A B)*(C D)	TABLES A*C A*D B*C B*D;
TABLES(A B C)*D	TABLES A*D B*D C*D;
TABLES (A--C)*D	TABLES A*D B*D C*D;
TABLES(A--D);	TABLES A B C D;

若未定义 TABLES 语句，则 FREQ 过程对数据集中的每个变量都生成一个一维频数表。在 FREQ 过程一次可包含多个 TABLES 语句。

TABLES 语句斜杠（/）后可用的主要控制选项如表 8-6 所示。

表 8-6　TABLES 语句后的可选项

选　项	意　义
CHISQ	要求对每一层的齐性或独立性进行 Λ^2 检验
MEASURES	要求一系列的线性关系指标和它们的标准误
JT	Jonckheere-Terpsta 检验
FISHER	FISHER'S 精确检验
BINOMAL	二项分布比率，置信度，对一维表格的检验
CMH	输出 Cochran-Mantel-Haenszel 相关统计量
ALL	要求 CHISQ、MEASURES、CMH 统计测试及计算关系指标
AGREE	要求计算且检验列联表中行变量和列变量的吻合程度
EXPECTED	在独立性或齐性的假设下，输出单元频数的期望值
DEVIATION	要求输出各单元期待频数和实际频数的差值
CELLCHI2	要求输出每一单元对 Λ^2 统计量的贡献
NOFREQ	不输出交叉表的单元频数
NOPERCENT	不输出交叉表的单元百分数
NOROW/NOCOL	不输出单元行或列百分数

续表

选　项	意　义
SCORE=RANK/TABLE/RIDIT/MODRIDIT	指明用何种数据执行 CMH 统计检验或计算皮尔逊相关系数。若 SCORE=TABLE，指分配表上行与列的次数，其他三种用来做非参数分析
ALPHA=	定义犯第一类错误的概率
MISSING	在计算百分数及其他统计量时包括缺失值
LIST	以非交叉表来表示频数结果
NOCUM	不输出累计频数和累计百分比

OUTPUT 语句——产生一个包含分析结果的输出数据文件，包括 TABLES 语句中定义的输出统计量、有效和遗漏数据的个数。此选项后可指定的关键字的字符串如表 8-7 所示。

表 8-7　关键字字符串

关 键 字	意　义
AJCHI	经过连续性校正后的 λ^2 值
ALL	由选项 CHISQ、CMH、MEASURE 导出的统计量的值和有效数据的个数
CHISQ	三个 λ^2 检验统计量的值
PLCORR	多元相关系数

注意：在界定包含在新数据集中的关键字前，读者必须在 TABLES 语句中定义其相关选项，否则此关键字的值将缺失。

8.2　拟合优度检验

8.2.1　基本原理

正确理解列联表的构成是进行列联分析的基础。在具体介绍列联表之前首先介绍本章两个重要的名词。

- 类别变量：类别变量取值为不连贯的数字。如名义变量 sex（取值为 Female、Male），次序变量 School（取值为 0——本科、1——硕士、2——博士），区间变量 Temperature（取值为 37℃～37.5℃、37.5℃～38℃），绝对变量 number（班级人数）。
- 类别数据：类别数据是来自多个观测在一个（或多个）类别变量上的取值。若类别数据来自观察体在一个类别变量上的数值时，可画出次数分配表；若类别数据来自观测在两个类别变量上的数据时，可用二维列联表表示，表格矩阵的行表示第一个变量所取的水平，列表示另一变量所取的水平；若类别数据来自两个以上的类别变量，用多维列联表或分观测表示。

列联表是两个或两个以上的类别变量交叉分组后形成的频数分布表，包括行变量、列变量及分层变量。一般行变量是分类变量，列变量是观测变量，如调查不同学历水平的人群对“是

否继续实行计划生育”持有的态度，行变量设置为学历分类，列变量设置为对这一论题持有的态度。

拟合优度检验主要用于检验类别间的频数是否满足一定的比例分布，如检验某地区的男女比是否为 1∶1 的步骤如下：

（1）提出假设：$H_0: n_1 : n_2 = 1$；$H_1: n_1 : n_2 \neq 1$。

（2）计算检验统计量：$\lambda^2 = \sum_{i=1}^{c} \frac{(f_i - e_i)^2}{e_i}$，其中 f_i（i 的取值为 1 和 2，分别代表男性和女性，则 c=2）代表列联表中第 i 列的实际频数，e_i 代表列联表中第 i 列的期望频数（即在原假设下计算出的频数，在本例中为总人数的一半），统计量的自由度为 $(c-1)=1$。

（3）下结论：根据显著性水平 α 和自由度查出临界值 $\lambda_\alpha^2(c-1)$，若 $\lambda^2 > \lambda_\alpha^2(c-1)$ 则拒绝原假设；若 $\lambda^2 < \lambda_\alpha^2(c-1)$，则接受原假设。

在实际应用中可将两类推广到多类，将检验 1∶1 比例推广到检验特定的比例分布，本实验将应用 SAS 的 FREQ 过程进行特定比例的拟合优度检验。

8.2.2 SAS 实例——检验各年龄阶层人口数是否满足特定分布

例 8-1 已知第五次人口普查不同年龄阶段的人口分布比和第六次人口普查不同年龄阶层的人口分布（如表 8-8 所示），试检验第六次人口普查不同年龄阶层人口分布是否与第五次人口普查一致？

表 8-8 人口普查年龄段分布数据

年龄阶段	A：0～14 岁	B：15～59 岁	C：60～65 岁	D：65 岁及以上
第五次普查人口分布比	22.9%	66.63%	2.8%	7.09%
第六次普查人口分布数	222459737	939616410	58816996	118831709

编写程序如下所示（其在光盘中的存储路径为“proc\chap8\population”）：

```
data chap8.population;   /*新建包含第六次人口普查不同年龄阶层的人口分布数据集*/
input age$ dis@@;
cards;
A 222459737      B 939616410      C 58816996       D 118831709
;
run;

 proc freq data=chap8.population order=data;
 /*调用 freq 过程，定义类别按变量值进入到数据集中的顺序排列*/
tables age / nocum chisq testp=(22.9 66.63 2.8 7.09);
/*对变量 age 进行一维拟合优度检验，定义检验比例为 22.9%：66.63%：2.8%：7.09%；
不输出累计频数和百分比*/
weight dis;                /*定义加权变量*/
run;
```

选择 Run|Submit 命令提交程序，以下分析输出结果：表 8-9 为第六次人口普查后各年龄阶段的人口出现频率（Frequency）、分布百分比（Percent）及需要检验的百分比（Test Percent）。

表 8-10 为对第六次人口普查不同年龄阶段人口分布是否满足第五次人口普查分布百分比的λ^2检验结果，检验 P 值小于 0.0001，则拒绝原假设，即第六次和第五次人口普查人群年龄分布不同。

表 8-9　一维分布表

age	Frequency	Percent	Test Percent
A	2.2246E8	16.60	22.90
B	9.3962E8	70.14	66.63
C	58816996	4.39	2.80
D	1.1883E8	8.87	7.09

表 8-10　卡方分布检验结果

Chi-Square Test for Specified Proportions	
Chi-Square	43739980.3599
DF	3
Pr > ChiSq	<.0001

8.3　独立性检验

8.3.1　基本原理

对二维列联表的行、列类别变量主要分析其独立性和相关度，独立性主要用λ^2 检验，一般的分析步骤如下。

（1）提出假设：H_0：行、列类别变量独立；H_1：行、列类别变量不独立。

（2）计算检验统计量：$\lambda^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(f_{ij}-e_{ij})^2}{e_{ij}}$，公式中的 f_{ij} 代表列联表中第 i 行、第 j 列类别的实际频数，e_{ij} 代表列联表中第 i 行、第 j 列类别的期望频数（由行、列变量独立的前提计算得到），自由度为$(r-1)(c-1)$。

（3）进行决策：根据显著性水平α和自由度$(r-1)(c-1)$查出临界值 λ_α^2，若 $\lambda^2>\lambda_\alpha^2$，则拒绝原假设；若 $\lambda^2<\lambda_\alpha^2$，则接受原假设。

在 SAS 系统中可应用 FREQ 过程以编程的方式和 Analyst 模块以菜单操作的方式完成独立性检验。

8.3.2　SAS 实例——居住地与驾车类型关系探索

例 8-2　某市场调查公司想了解某地的城市成年人驾驶汽车的类型与他们居住地之间是否存在一定的联系，特随机抽取了 546 名成年驾驶者进行调查，得到结果如表 8-11 所示（相应的

SAS 数据集在光盘中的存储路径为“data\chap8\car_survey”)，请问驾车类型和居住地是否独立？

表 8-11　居住地与驾车类型调查数据

居　住　区	驾 车 类 型		
	A	B	C
1	52	64	26
2	69	63	32
3	50	45	28
4	45	53	19

编程法：

编写程序如下所示：

```
proc freq data=chap8.car_survey;                    /*调用 freq 过程*/
tables area*car_type/chisq;
/*定义列联表行变量为 area，列变量为 car_type，并输出卡方检验结果*/
weight number;                                      /*定义加权变量为 number*/
run;
```

选择 Run|Submit 命令提交程序，以下分析输出结果：表 8-12 为行变量为 area、列变量为 car_type 的二维列联表。每个单元格中从上至下依次为对应类型的频数、总百分比、行百分比和列百分比。表 8-13 为独立性检验结果和衡量相关性的统计量，由于独立性检验的λ^2检验、似然比λ^2（likelihood Ratio Chi-Square）、Mantel-Haesszel λ^2检验对应的检验 P 值分别为 0.6674、0.6785、0.8198，大于显著性水平 0.05，则接受原假设，认为变量 area 和变量 car_type 是独立的，即人群驾驶的车辆类型分布在不同的地区分布没有显著的差异。同时 Phi 系数（Phi Coefficient）、因变系数（Contingency Coefficient）和克拉默值（Cramer's V）这些从皮尔逊λ^2系数公式中衍生出来衡量相关关系的统计量的值都在 0.08 左右，同样说明变量 area 和 car_type 是独立的。

表 8-12　二维列联表

Table of area by car_type				
area(area)	car_type(car_type)			
Frequency Percent Row Pct Col Pct	A	B	C	Total
1	52 9.52 36.62 24.07	64 11.72 45.07 28.44	26 4.76 18.31 24.76	142 26.01
2	69 12.64 42.07 31.94	63 11.54 38.41 28.00	32 5.86 19.51 30.48	164 30.04
3	50 9.16 40.65 23.15	45 8.24 36.59 20.00	28 5.13 22.76 26.67	123 22.53

续表

Table of area by car_type				
area(area)	car_type(car_type)			
Frequency Percent Row Pct Col Pct	A	B	C	Total
4	45 8.24 38.46 20.83	53 9.71 45.30 23.56	19 3.48 16.24 18.10	117 21.43
Total	216 39.56	225 41.21	105 19.23	546 100.00

表 8-13　λ^2 检验结果

Statistic	DF	Value	Prob
Chi-Square	6	3.9948	0.6774
Likelihood Ratio Chi-Square	6	3.9865	0.6785
Mantel-Haenszel Chi-Square	1	0.0519	0.8198
Phi Coefficient		0.0855	
Contingency Coefficient		0.0852	
Cramer's V		0.0605	

菜单法：

步骤一：选择 Solutions|Analysis|Analyst 命令，进入 Analyst 分析界面。

步骤二：选择 File|Open By SAS Name|chap8|car_survey 命令，打开数据集 chap6.car_survey。

步骤三：选择 Statistics|Table Analysis 命令，弹出如图 8-1 所示对话框，单击变量 area，再单击 Row（行）按钮，将变量 area 选为行变量，用同样的方法将变量 car_type 选为列变量（Column），将变量 number 选进 Cell Counts（单元格计数）。

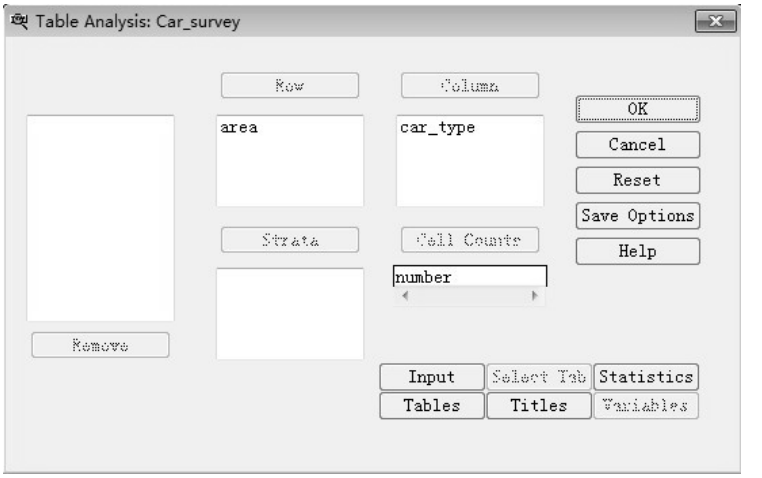


图 8-1　列联表分析主对话框

单击 Statistics（统计量）按钮，弹出如图 8-2 所示对话框，选择 Chi-square statistics，定义输出卡方统计量。若不希望输出频数表，可勾选表格下方的 Print statistics only（仅输出统计量）；若希望计算时不删除缺失值，可勾选 Include missing values in calculations（计算时包含缺失值）。单击 OK 按钮保存设置，并返回如图 8-1 所示对话框。单击 OK 按钮按钮则输出和编程方法一致的检验结果。

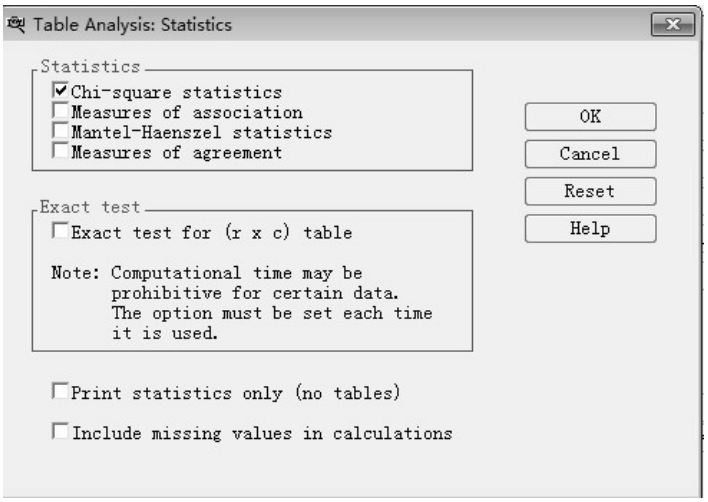


图 8-2　选择输出统计量

8.4　一致性检验

8.4.1　基本原理

一致性检验主要用于配对样本的比较，目的是通过单一样本数据推断两种处理结果是否一致。此检验方法在医学统计学中应用较多。例如，研究者希望考察不同的诊断方法在诊断结果上是否具有一致性，如评价两种诊断试验方法对同一个样本（化验对象）的化验结果的一致性、两个医务工作者对同一组病人的诊断结论的一致性、同一医务工作者对同一组病人前后进行两次观察做出诊断的一致性等。1960 年 Cohen 等人提出用 Kappa 统计量作为评价判断一致性程度的指标。实践证明，它是一个描述一致性较为理想的指标，因此现在得到广泛的应用。Kappa 系数作为评价判断的一致性程度的重要指标，取值在 0～1 之间。一般情况下，Kappa≥0.75，则两者一致性较好；0.75>Kappa≥0.4，则两者一致性一般；Kappa<0.4，则两者一致性较差。

8.4.2　SAS 实例——HR 对求职者评定等级的一致性研究

例 8-4　为了让公司的面试更加科学化和规范化，人力资源管理部门在某一面试环节分别让两名面试官对相同的求职者进行面试，并且划分评定等级，记录结果如表 8-14 所示，相应的 SAS

数据集在光盘中的存储路径为“data\chap8\HR”。试问这两名面试官的评定标准是否一致？

表 8-14　面试评定等级数据

HR1	HR2				
	A	B	C	D	E
A	20	14	11	10	8
B	15	21	22	12	7
C	12	14	22	15	5
D	8	2	6	13	4
E	3	1	2	4	19

编程法：

编写如下程序（其在光盘中的存储路径为“proc\chap8\hr”）：

```
proc freq data=chap8.hr order=data;
/*调用 freq 过程，且变量取值按进入数据集的先后排序*/
tables hr1*hr2 / agree;
/*定义二维列联表的行变量为 hr1，列变量为 hr2，进行一致性检验*/
test Kappa;                                    /*检验 Kappa 统计量*/
weight count;                                  /*定义加权变量为 count*/
run;
```

选择 Run|Submit 命令提交程序，以下分析主要输出结果：由表 8-15 可知 Kappa 系数为 0.1843，它的 95%置信区间为（0.1119，0.2567），则两个面试官的评价一致性较差。表 8-16 为加权 Kappa 系数的值（0.2532）、它的 ASE 检验 *P* 值及 95%置信区间。这一结果进一步验证了两个面试官对应聘者的评价的一致性较差。

表 8-15　Kappa 系数

Simple Kappa Coefficient	
Kappa	0.1843
ASE	0.0369
95% Lower Conf Limit	0.1119
95% Upper Conf Limit	0.2567

表 8-16　加权 Kappa 系数

Weighted Kappa Coefficient	
Weighted Kappa	0.2532
ASE	0.0448
95% Lower Conf Limit	0.1654
95% Upper Conf Limit	0.3411

菜单法：

步骤一：选择 Solutions|Analysis|Analyst 命令，进入 Analyst 分析界面。

步骤二：选择 File|Open By SAS Name|chap8|hr 命令，打开数据集 chap8.hr。

步骤三：选择 Statistics|Table Analysis 命令，弹出如图 8-3 所示对话框，单击变量 hr1，再单击 Row（行）按钮，则将变量 hr1 选为行变量，用同样的方法将变量 hr2 选为列变量，将变量 count 选进 Cell Counts（单元格计数）。

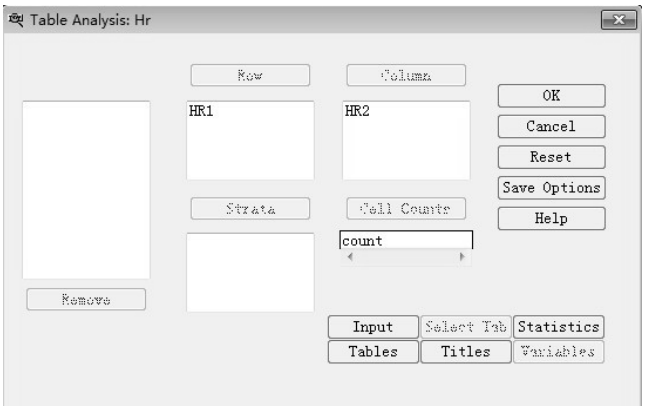


图 8-3　列联表分析主对话框

单击 Input（进入）按钮，弹出如图 8-4 所示对话框，单击选择 Order of appearance in data set（变量取值按其进入数据集的先后排序）。单击 OK 按钮保存设置并返回如图 8-3 所示对话框。

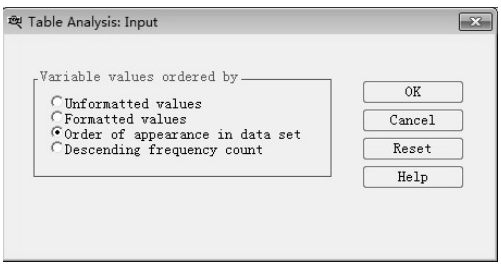


图 8-4　设置变量输入顺序

单击 Statistics（统计量）按钮，弹出如图 8-5 所示对话框，选择 Measures of agreement（一致性度量），定义输出一致性度量统计量。单击 OK 按钮保存设置并返回如图 8-3 所示对话框。单击 OK 按钮，则输出与编程法类似的结果。

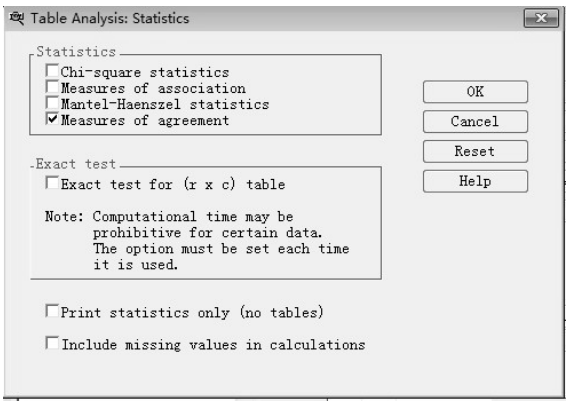


图 8-5　选择统计量

8.5　属性关联度

8.5.1　基本原理

有序变量的属性关联分析可用 Spearman 秩相关系数等统计量来衡量，Spearman 秩相关系数的计算过程为：首先找出所有 X_i 在 X 样本中的秩 R_i（即将所有样本从小到大排序，X_i 的秩为它的排列顺序），用同样的方式找出所有 Y_i 在 Y 样本中的秩 S_i，记 $\overline{R}=\frac{1}{n}\sum_{i=1}^{n}R_i$，$\overline{S}=\frac{1}{n}\sum_{i=1}^{n}S_i$，Spearman 相关系数定义如下：

$$r_s=\frac{\sum_{i=1}^{n}(R_i-\overline{R})(S_i-\overline{S})}{\sum_{i=1}^{n}(R_i-\overline{R})^2\sum_{i=1}^{n}(S_i-\overline{S})^2}$$

在 SAS 系统中可用 FREQ 过程以编程的方式和 Analyst 模块以菜单操作的方式计算 Spearman 相关系数。

8.5.2　SAS 实例——探索某原料的产地与质量等级的关系

例 8-5　一种原料来自三个不同规模的原产地，原料质量被分成三个不同等级（A、B、C 依次递减）。从这批原料中随机抽取 677 件进行检验，结果如表 8-17 所示，相应的 SAS 数据集在光盘中的存储路径为“data\chap8\quality_level”。试问地区和原料之间是否存在依赖关系（α=0.05）？

表 8-17　产品质量等级抽查结果

地　区	A 级	B 级	C 级
大	89	68	34
中	50	49	42
小	100	165	80

编程法：

编写程序如下所示（其在光盘中的存储路径为“proc\chap8\quality_level”）：

```
proc freq data=chap8.quality_level;          /*调用 freq 过程*/
table area*quality_level/chisq measures;     /*定义列联表行变量为 area,列变量为 quality_level；
                                               输出卡方统计量及属性变量相关系数*/
weight number;                               /*定义加权变量为 number*/
run;
```

选择 Run|Submit 命令提交程序，以下分析主要输出结果：表 8-18 为独立性检验结果，三个卡方检验对应的 P 值都小于 0.05，则拒绝原假设，认为变量 area 和 quality_level 之间不独立。而衡量相关性的 Phi 和 Contingency 系数在 1.8 左右，Cramer's V 取值约为 0.13，说明原料产地

和它的质量之间存在一定相关性。表 8-19 为属性关联分析统计量，Statistic 列为所有的统计量、Value 列为相应统计量的值，ASE 列为备择假设为“该统计量的值为零”假设检验的 *P* 值。因本实验采用的是定序数据，则用 Spearman、Kendall's Tau-b 系数来测定属性关联度，Spearman 系数值为 0.1171，Kendall's Tau-b 系数取值为 0.1068。即变量 area 和 quality_level 存在正相关，则可判断出随着原产地规模的减小，产品的质量也是随之下降的。

表 8-18　检验统计量

Statistic	DF	Value	Prob
Chi-Square	4	22.3135	0.0002
Likelihood Ratio Chi-Square	4	21.9304	0.0002
Mantel-Haenszel Chi-Square	1	8.9374	0.0028
Phi Coefficient		0.1815	
Contingency Coefficient		0.1786	
Cramer's V		0.1284	

表 8-19　属性关联分析统计量

Statistic	Value	ASE
Gamma	0.1665	0.0560
Kendall's Tau-b	0.1068	0.0362
Stuart's Tau-c	0.1014	0.0344
Somers' D C|R	0.1094	0.0371
Somers' D R|C	0.1041	0.0353
Pearson Correlation	0.1150	0.0396
Spearman Correlation	0.1171	0.0397
Lambda Asymmetric C|R	0.0557	0.0394
Lambda Asymmetric R|C	0.0000	0.0000
Lambda Symmetric	0.0303	0.0216
Uncertainty Coefficient C|R	0.0151	0.0065
Uncertainty Coefficient R|C	0.0158	0.0067
Uncertainty Coefficient Symmetric	0.0154	0.0066

菜单法：

步骤一：选择 Solutions|Analysis|Analyst 命令，进入 Analyst 分析界面。

步骤二：选择 File|Open By SAS Name|chap8|quality_level 命令，打开数据集 chap8. quality_level。

步骤三：选择 Statistics|Table Analysis 命令，弹出如图 8-6 所示对话框。，单击变量 area，再单击 Row（行）按钮，则将变量 area 选为行变量，用同样的方法将变量 quality_level 选为列变量，将变量 number 选进 Cell Counts（单元格计数）。

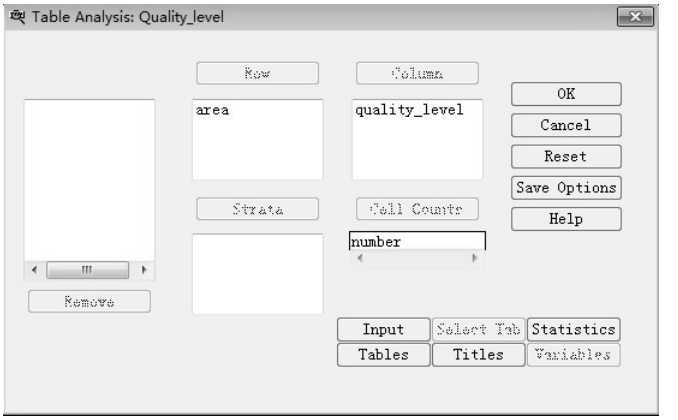


图 8-6　列联表分析主对话框

单击 Statistics（统计量）按钮，弹出如图 8-7 所示对话框，选择 Chi-square statistics（卡方统计量）、Measures of association（属性相关度量），即设定进行属性变量的独立性检验，输出属性相关统计量。单击 OK 按钮保存设置并返回如图 8-6 所示对话框。单击 OK 按钮则将输出与编程法一致的结果。

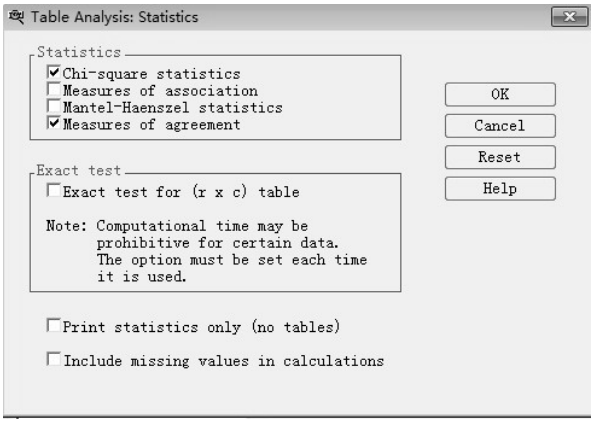


图 8-7　选择输出统计量

练习题

习题 8-1　已知 2000 年某学院的教职工学历构成比和 2010 年不同学历的教师数目（如表 8-20 所示），试检验经过 10 年该学院的教师学历构成是否发生了显著的变化？（α=0.05）

表 8-20　教职工学历构成

年龄阶段	博士	硕士	本科	本科以下
2000 年教职工学历构成比	15.3%	42.6%	24.5%	17.6%
2010 年教职工学历构成数	35	59	42	11

（本习题的解答程序在光盘中的存储路径为“proc\chap8\teacher”。）

习题 8-2 为探索成年男性饮酒和抽烟的习惯是否独立，调查了 100 位男性，询问他们是否有饮酒或吸烟的习惯，得到了频数分布表如表 8-21 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap8\habit”）。请选择合适的分析方法并得出结论。

表 8-21 抽烟和饮酒

抽烟（smoking）	喝酒（drink）	
	是（1）	否（0）
是(1)	42	10
否(0)	14	34

（本习题的解答程序在光盘中的存储路径为“proc\chap8\habit”。）

习题 8-3 某电视事业推广部为了解 18～35 岁和 35 岁以上的观众对新推出的某一电视节目的评价是否具有一致性，特进行了市场调查，数据如表 8-22 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap8\TV_survey”）。请选择合适的分析方法并得出结论。

表 8-22 节目调查数据

35 岁以上（age2）	18～35 岁（age1）			
	非常喜欢	喜欢	一般	不喜欢
非常喜欢	35	32	14	16
喜欢	23	26	11	24
一般	12	27	12	15
不喜欢	13	12	26	11

（本习题的解答程序在光盘中的存储路径为“proc\chap8\TV_survey”。）

习题 8-4 为了探索年级和学生成绩等级之间的关系，某人从学院教务处查得某班学生从大一至大四的不同成绩等级的学生数目，如表 8-23 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap8\level”）。请选择合适的分析方法并得出相应的结论。

表 8-23 不同年级的学生成绩等级分布情况

年级（grade）	学生成绩等级（level）				
	A	B	C	D	E
大一	12	24	19	5	7
大二	10	23	21	7	5
大三	5	20	25	9	8
大四	4	15	30	12	7

（本习题的解答程序在光盘中的存储路径为“proc\chap8\level”。）

第 9 章　非参数检验

在统计推断和假设检验中，依赖于确定的、包含自由参数的概率分布的检验称为参数检验。参数检验通常对数据分布有一定假定条件，而在实际测量中，某些实验结果只有程度上的区别，如颜色深浅、喜好强度等，此类数据通常不满足特定分布，因此也不再适用于参数检验，此时可考虑应用本章介绍的非参数统计分析方法（Nonparametric Statistics）。该方法对数据总体分布仅做极少的假设甚至无须任何假设，而尽量从数据本身获取需要的信息。非参数检验主要应用了数据符号和秩（rank）的信息，若把一组数据按大小次序排队，每个数据在整组数据中（从最小的数起）的位置和次序被称为该数据的秩。非参数检验的具体内容在《非参数统计》（吴喜之编，中国统计出版社）中有详细而清晰的介绍。

本章首先介绍单因子非参数方差分析 NPAR1WAY 过程，再具体介绍单样本位置检验、配对的和独立的两样本位置检验及多个独立样本的方差分析的基本原理、编程分析和菜单分析方法。

9.1　SAS 过程——NPAR1WAY 过程

NPAR1WAY 过程主要计算基于经验分布的函数（EDF）的和通过一个单因素分类变量的响应变量确定的秩的得分（具体包括 WILCOXON 得分、中位数得分、SAVAGE 得分和 VAN DER WAERDEN 得分），再由秩得分计算简单的线性秩统计量，以此检验一个变量的分布在不同组中是否具有相同的位置参数。

NPAR1WAY 过程的一般使用格式为：

```
PROC NPAR1WAY   DATA=数据集 <选项列表>;
CLASS     分类变量;
VAR       变量列表;
BY        变量列表;
RUN;
```

PROC NPAR1WAY 和 CLASS 语句在使用时必须定义，其他语句根据需要选择。PROC NPAR1WAY 语句后的主要可选控制项如表 9-1 所示。

表 9-1　PROC NPAR1WAY 语句后的主要控制选项

选　项	意　义
ANOVA	对原始数据执行标准方差分析
EDF	计算基于经验分布函数（EDF）的统计量，如 KOLMOGOROV-SMIRNOV、CRAMER-VON MESES、KUIPER 统计量
MISSING	把 CLASS 变量的缺失值看作一个分类水平
MEDIAN	执行一个中位数得分分析。对于两样本产生一个中位数检验，对于更多样本产生一个 BROWN-MOOD 检验

续表

选　项	意　义
SAVAGE	执行一个 SAVAGE 得分分析。该检验适用于数据服从指数分布的组间比较
VW	执行一个 VAN DER WAERDEN 得分分析。这是一个通过应用反正态分布累积函数得到近似的正态得分。对于两个水平情况，这是一个标准 VAN DER WAERDEN 检验
WILCOXON	对数据或 WILCOXON 得分进行秩分布。对于两个水平，它与 WILCOXON 秩和检验一样；对于任何数量的水平，这是一个 KRUSKAL-WALLIS 检验
ST	执行两样本的 SIEGEL-TUKEY 方差检验

NPAR1WAY 过程中主要使用的语句含义如下：

CLASS 语句——指定一个（有且只能有一个）分类变量用来标识数据中的各个类。CLASS 语句指定的变量可为字符型或数值型。

VAR 语句——指定分析变量，若省略 VAR 语句，NPAR1WAY 过程将分析数据集中除语句指定的变量外的所有数值型变量。

BY 语句——定义分组分析变量，使用时要求将数据集预先按 BY 变量排序。

9.2　单样本位置检验

9.2.1　基本原理

单样本位置检验的目的是检验某个样本均值是否等于特定值，最简单的非参数单样本位置检验是符号检验（Sign Test）。它适用于样本和总体中位数的比较、数据的升降趋势的检验等。

符号检验的基本思路为：首先定义样本中数据取值符号（正号或负号）的规则，如检验某个样本均值是否等于特定值，则将样本中每一个数与检验值比较，大于比值的定义为正号，小于此值的定义为负号，而等于特定值的样本被删除。计数正号的个数 S^+ 及负号的个数 S^-，此时样本量为 $n=S^++S^-$。当样本 n 较小时，应使用二项分布概率计算方法计算检验 P 值；当样本 n 较大时，常利用二项分布的正态近似计算检验 P 值，以下分别介绍。

1．小样本时的二项分布概率计算

当 $n\leqslant 20$ 时，S^+ 或 S^- 的检验 P 值由精确计算尺度二项分布的卷积获得。以下分三种情形讨论：

- 当原假设为“样本均值和特定值相等，即出现正负号的个数相等，即 $S^+=S^-$”，则正号出现的概率 $P=0.5$，于是 S^+ 与 S^- 均服从二项分布 $B(n,0.5)$，对于太大的 S^+ 而相应太小的 S^-，或者太大的 S^- 而相应太小的 S^+，都将拒绝原假设。此时选择 $\min(S^+,S^-)$ 作为检验统计量，检验 $P_{检}$：

$$P_{检}=2P(K\leqslant k)=2\sum_{i=0}^{k}C_n^i(0.5)^i(1-0.5)^{(n-i)}=\frac{1}{2^{n-1}}\sum_{i=0}^{k}C_n^i$$

- 当原假设为“样本均值大于特定值”，正号的个数 S^+ 大于负号的个数 S^- 的可能性应该大，

即正号出现的概率 $P>0.5$，对于太小的 S^+ 而相应太大的 S^-，将拒绝原假设。选择 $\min(S^+,S^-)$ 作为检验统计量，检验 $P_{检}$：

$$P_{检}=P(K\leqslant k)=\sum_{i=0}^{k}C_n^i(0.5)^i(1-0.5)^{(n-i)}=\frac{1}{2^n}\sum_{i=0}^{k}C_n^i$$

- 当原假设为“样本均值小于特定值”，正号的个数 S^+ 小于或等于负号的个数 S^- 的可能性应该大，即正号出现的概率 $p\leqslant 0.5$，对于太大的 S^+ 而相应太小的 S^-，将拒绝原假设。选择 $\min(S^+,S^-)$ 作为检验统计量，检验 $P_{检}$：

$$P_{检}=P(K\leqslant k)=\sum_{i=0}^{k}C_n^i(0.5)^i(1-0.5)^{(n-i)}=\frac{1}{2^n}\sum_{i=0}^{k}C_n^i$$

2．大样本时的正态近似概率计算

当样本量 $n>20$ 时，我们可以利用二项分布的正态近似，即对于 $S\sim B(n,P)$，二项分布的期望均值为 nP，方差为 $nP(1-P)$，当 n 比较大时，且 nP 和 $n(1-P)$ 大于 5，可近似地认为：

$$z=\frac{S-nP}{\sqrt{nP(1-P)}}\sim N(0,1)$$

公式中的 S 表示正号或负号的个数，符号检验时，P=0.5，此时得到大样本时的正态近似统计量：

$$z=\frac{S-0.5n}{0.5\sqrt{n}}\sim N(0,1)$$

当 $S>n/2$ 时，应该修正 S 为 S −0.5；当 $S<n/2$ 时，应该修正 S 为 S +0.5。在统计量中增加连续性修正因子 0.5 的目的是将连续分布应用到近似的离散型分布。

9.2.2　SAS 实例——检验某工地施工是否提高小区噪声水平

例 9-1　某住宅小区的夜间噪声一直保持在 30dB（分贝）。后来附近有建筑工地施工。表 9-2 是连续 20 天夜间在该小区测得的噪声水平（分贝）数据（相应的 SAS 数据集在光盘中的存储路径为“data\chap9\noise”）。请问该建筑工地是否提高了小区夜间噪声水平？（$\alpha=0.05$）

表 9-2　某小区噪声水平测量值　　单位：分贝

58	37	46	50	32	50	47	56	33	34
43	38	37	21	23	26	32	25	53	45

解析：本例的原假设为“该建筑工地没有提高小区夜间噪声水平”，备择假设为“该建筑工地显著地提高了此小区夜间的噪声水平”。

编写如下程序（其在光盘中的存储路径为“proc\chap9\noise”）：

```
proc univariate data=chap9.noise   mu0=30;   /*调用 univariate 过程，检验样本的中位数是否为 30*/
var noise;                  /*指定分析变量为 noise*/
run;
```

注意：SAS 的 univariate 过程只能进行双侧检验，但是由于无论双侧还是单侧检验，检验统计量相同，而且双侧检验对应的 P 值是单侧检验的两倍，因此本例由软件计算出的 P 值应该再除以 2 才是真正单侧检验的 P 值。

选择 Run|Submit 命令提交程序，以下分析主要输出结果：表 9-3 为样本的基本描述性统计量，包括提供样本位置信息的均值（Mean）、中位数（Median）和众数（Mode）及提供样本离散程度信息的标准差（Std Deviation）、方差（Variance）、极差（Range）和分位数极差（Interquartile Range）。表 9-4 为对原假设“样本中位数为 30（即施工后该地的噪声平均水平仍然保持在 30）”的检验结果，非参数检验符号检验（Sigh）统计量的值为 6，对应的双边检验的 P 值为 0.0118，则单边检验 P 值为 0.0059，小于显著性水平 0.05，则应该拒绝原假设，接受备择假设，认为该地由于施工导致噪声明显上升，要求物业管理和社会治安部门予以干涉。

表 9-3　样本基本描述性统计指标

Basic Statistical Measures			
Location		Variability	
Mean	39.30000	Std Deviation	11.15489
Median	37.50000	Variance	124.43158
Mode	32.00000	Range	37.00000
		Interquartile Range	16.50000

表 9-4　位置检验结果

Tests for Location: Mu0=30				
Test	Statistic		p Value	
Student's t	t	3.728487	Pr > |t|	0.0014
Sign	M	6	Pr >= |M|	0.0118
Signed Rank	S	75.5	Pr >= |S|	0.0032

9.3　Wilcoxon 符号秩检验

9.3.1　基本原理

9.2 节讲解的符号检验利用了观察值和特定值的位置之差的符号来进行检验，但是它没用充分利用这些差值的大小所包含的信息。不同符号代表在特定值的哪一侧，而差的绝对值的秩的大小代表了距离特定值的远近。因此如果把两者结合起来检验的效率应该更高，即为本节介绍的 Wilcoxon 符号秩检验（Wilcoxon sign rank test）。本检验的目的和 Wilcoxon 检验是一致的。本检验在样总体分布满足对称性和连续性假定时适用。检验原假设某样本中位数为特定值 $\mathrm{mode}(X)=M$ 。

检验步骤为：

（1）对样本中的每一个观测值 $X_i(i=1,\cdots,n)$，计算 $|X_i-M|$，代表这些样本点到 M 的距离。

（2）把以上计算出的 n 个绝对值排序，得到对应的秩，如果出现打结的情形（即有相同的样本点，则对每个点取平均值（如 1，2，2，4 对应的秩为 1，2.5，2.5，4）。

（3）令W^+等于$X_i-M>0$的$|X_i-M|$的秩的和。类似的，W^-等于$X_i-M<0$的$|X_i-M|$的秩的和。注意到$W^++W^-=n(n+1)/2$。

（4）对双边检验，即备择假设$\mathrm{mode}(X)\neq M$时，在原假设下，W^+和W^-的值应该差不多，因此如果其中之一很小时，将怀疑原假设；此时选择检验统计量为$\min(W^+,W^-)$。类似的，对$H_0:\mathrm{mode}(X)\leqslant M;H_1:\mathrm{mode}(X)>M$的单边检验取统计量为$W^-$，对$H_0:\mathrm{mode}(X)\geqslant M$；$H_1:\mathrm{mode}(X)<M$的单边检验取统计量为$W^+$。

（5）根据得到的统计量通过查 Wilcoxon 符号秩的分布表得到原假设下的检验 P 值，在大样本情形下用正态近似，得到一个与 W 有关的正态随机变量 Z：

$$Z=\frac{W-n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}\sim N(0,1)$$

再计算出检验 P 值。最后将检验 P 值和设定的显著性水平α进行比较，如果 P 值小于α，则拒绝原假设；否则将接受原假设。

9.3.2　SAS 实例——情绪对血压值的影响

例 9-2　某研究小组抽取了 16 个志愿者分别测得他们在情绪低落和高涨时的血压值（两次测量时，其他外在条件一致），并计算出平均血压，具体数值如表 9-5 所示，相应的 SAS 数据集在光盘中的存储路径为“data\chap9\emontion”。试问情绪的变化是否会导致平均血压值的变化？

表 9-5　情绪变动时测量得到血压平均值　　单位：mmHg

情绪低落时	110	123	105	122	118	114	120	107
情绪高涨时	114	127	101	128	123	120	127	119
情绪低落时	109	113	108	121	122	120	115	117
情绪高涨时	105	116	111	122	128	125	110	121

编写程序如下所示（其在光盘中的存储路径为“data\chap9\emontion”）：

```
/*新建临时文件 temp,变量 diff 为受试者情绪高涨时和情绪低落时平均血压的差值*/
data temp;
set chap9.emontion;
diff=high-low;
run;
proc univariate data=temp mu0=0; /*调用 univariate 过程*/
var diff;                        /*指定分析变量为 diff*/
run;
```

选择 Run|Submit 命令提交程序，以下分析主要输出结果：表 9-6 为样本的基本描述性统计量，观察此表可知样本 diff 的均值为 3.3125，中位数和众数为 4。表 9-7 为对原假设“样本均值为零”的检验结果，非参数检验符号检验（Sigh）统计量的值为 5，对应的双边检验的值为 0.0213，小于显著性水平 0.05，非参数符号秩检验（Signed Rank）统计量的值为 46，对应的双边检验 P 值为 0.0147，同样小于显著性水平 0.05，则拒绝原假设。考虑本实验的背景，认为人在情绪低落和高涨时血压平均值有显著的差异，即情绪可能影响血压水平。

表 9-6　样本基本描述性统计指标

Basic Statistical Measures			
Location		Variability	
Mean	3.312500	Std Deviation	4.46794
Median	4.000000	Variance	19.96250
Mode	4.000000	Range	17.00000
		Interquartile Range	4.00000

表 9-7　位置检验结果

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	2.965572	Pr > |t|	0.0096
Sign	M	5	Pr >= |M|	0.0213
Signed Rank	S	46	Pr >= |S|	0.0147

9.4　Wilcoxon 秩和检验

9.4.1　基本原理

当两个独立样本来自正态分布和具有相同方差时，一般采用 T 检验比较均值。当样本并不满足这两个条件时可以考虑采用 Wilcoxon 秩和检验。Wilcoxon 秩和检验首先将两样本看成是单一样本，由小到大排列观察值统一编秩。如果原假设两个独立样本来自相同的总体为真，那么秩将大约均匀分布在两个样本中。如果备择假设两个独立样本来自不相同的总体为真，那么其中一个样本将会有较多的小秩值，从而得到一个较小的秩和；相应的也会得到一个较大的秩和。

设两个独立样本为：第一个 x 的样本容量为 n_1，第二个 y 样本容量为 n_2，在容量为 $n=n_1+n_2$ 的混合样本（第一个和第二个）中，x 样本的秩和为 W_x，y 样本的秩和为 W_y，再由 $\min(W_x,W_y)$ 满足样本大小为 n_1 和 n_2 的 Mann-Whitney-Wilcoxon 分布，得到检验 P 值。

在实际应用中，对于每个样本中的观察数大于或等于 8 的大样本，可采用标准正态分布 z 来近似检验。推导出标准化 W_x 为：

$$z=\frac{W_x-\mu\pm0.5}{\sigma}=\frac{W_x-\frac{n_1(n_1+n_2+1)}{2}\pm0.5}{\sqrt{\frac{n_1n_2(n_1+n_2+1)}{12}-\frac{n_1n_2\sum(\tau_j^3-\tau)}{12(n_1+n_2)(n_1+n_2-1)}}}\sim N(0,1)$$

其中分子±0.5 对离散变量进行连续性修正，对于 $W_x-\mu$ 大于 0−0.5 修正，对于 $W_x-\mu$ 小于 0+0.5 修正。然后再计算出显著性检验 P 值。

9.4.2　SAS 实例——检验两地地表土壤的 pH 值的差异

例 9-3　两个地点的地表土壤的 pH 值测量值如表 9-8 所示，相应的 SAS 数据集在光盘中的存储路径为“data\chap9\ph”。请问这两个地点的平均 pH 值是否一样？

表 9-8　地表 PH 值测量数据

地点 A	8.12	8.23	8.01	7.99	7.93	7.84	7.83	7.86	7.80
地点 B	7.88	7.90	8.02	7.59	7.78	7.69	7.50	7.37	7.23

编程法：

编写程序如下所示（其在光盘中的存储路径为“proc\chap9\ph”）：

```
proc npar1way data=chap9.ph wilcoxon ;      /*调用 npar1way 过程，指定 Wilcoxon 秩和检验*/
  class site;                /*定义分组变量为 site*/
  var ph_value;              /*定义分析变量为 ph_value*/
  run;
```

选择 Run|Submit 命令提交程序，以下分析主要输出结果：表 9-9 为变量 ph_value 根据变量 site 分组的得分情况：地点 A 对应的 pH 值的 Wilcoxon 总得分为 111，平均得分约为 12.33，在原假设“两个地点土壤的 pH 值相同”的条件下地点 A 对应的 pH 值期望总 Wilcoxon 得分为 85.5，期望得分标准差为 11.324752。地点 B 的 pH 值 Wilcoxon 得分情况类似。

表 9-9　不同分组的 Wilcoxon 得分

Wilcoxon Scores (Rank Sums) for Variable ph_value Classified by Variable site					
site	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
A	9	111.0	85.50	11.324752	12.333333
B	9	60.0	85.50	11.324752	6.666667

表 9-10 为两样本 Wilcoxon 秩和检验的结果，检验统计量为 111，正态近似的 Z 检验统计量等于 2.2076，对应的左侧检验和双侧检验的 *P* 值都小于显著性水平 0.05；*t* 分布近似对应的检验左侧和双边检验的 *P* 值也都小于 0.05，则拒绝原假设，认为两地土壤的酸碱性不一致。表 9-11 为两样本的 Kruskal-Wallis 检验结果，Chi-Square 检验统计量为 5.0702，自由度为 1，对应的检验 *P* 值为 0.0243，小于显著性水平 0.05，则拒绝原假设，进一步支持了 Wilcoxon 秩和检验结果。

表 9-10　Wilcoxon 两样本检验结果

Wilcoxon Two-Sample Test	
Statistic	111.0000

续表

Wilcoxon Two-Sample Test	
Normal Approximation	
Z	2.2076
One-Sided Pr > Z	0.0136
Two-Sided Pr > |Z|	0.0273
t Approximation	
One-Sided Pr > Z	0.0207
Two-Sided Pr > |Z|	0.0413
Z includes a continuity correction of 0.5.	

表 9-11　Kruskal-Wallis 检验结果

Kruskal-Wallis Test	
Chi-Square	5.0702
DF	1
Pr > Chi-Square	0.0243

菜单法：

步骤一：选择 Solutions|Analysis|Analyst 命令，进入 Analyst 分析主界面。

步骤二：选择 File|Open By SAS Name|chap9|ph|OK 命令，打开数据集 chap11.ph。

步骤三：选择 Statistics|ANOVA|Nonparametric One-Way ANOVA 命令，弹出非参数方差分析对话框，如图 9-1 所示。单击变量 ph_value，再单击 Dependent（因变量）按钮，则将变量 ph_value 选为因变量。类似的，将变量 site 选进 Independent（自变量）选项框内。

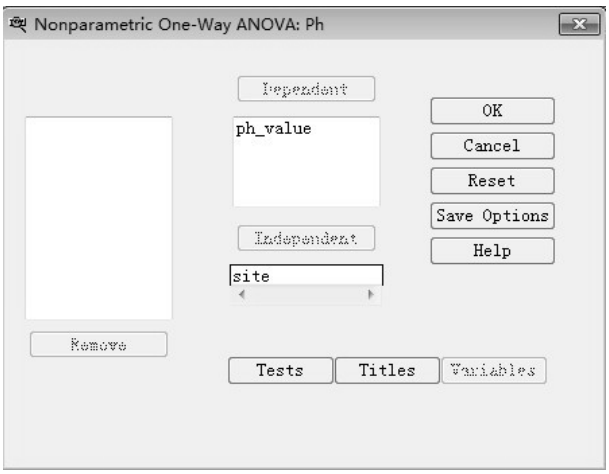


图 9-1　单因素非参数方差分析对话框

单击 Tests（检验）按钮，弹出如图 9-2 所示对话框，系统默认选择 Location test scores（位置得分检验）下的 Wilcoxon（Kruskal-Wallis test），即设置应用 Wilcoxon（Kruskal-Wallis test）

检验不同地点（site）土壤的酸碱性（ph_level）是否一致。在 Dispersion test scores（尺度得分检验）下提供了 4 种尺度检验的方法，即检验两样本的离散程度是否一致。在 Exact p-values（计算精确 *P* 值）下可设置计算精确的检验 *P* 值，在数据量较大时选择此项会延长系统运算的时间，且此选项每次使用都需重新设置。单击 Cancel 按钮采用系统默认设置并返回如图 9-1 所示对话框。单击 OK 按钮则将输出与编程方法一致的检验结果。

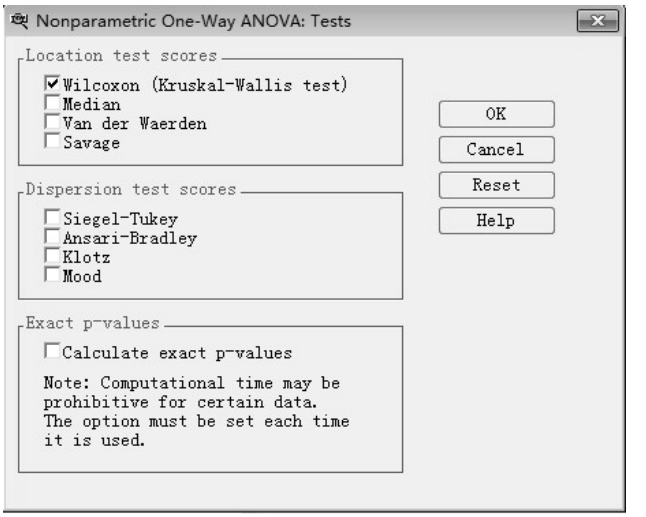


图 9-2　设置检验方法

9.5　Kruskal-Wallis 秩和检验

9.5.1　基本原理

当多个样本数据满足正态分布且具有相等的方差时，比较它们的均值可以采用方差分析。当数据不满足上述假定条件时，可考虑采用多样本均值比较的非参数的 Kruskal-Wallis 秩和检验，本质上它是两样本 Wilcoxon 秩和检验方法的推广。

Kruskal-Wallis 秩和检验，首先将所有样本的值混合在一起看成是单一样本，再把这个单一的混合样本中的值从小到大排序，序列值替换成秩值，最小的值给予秩值 1，遇到打结的情形则平分秩值。将数据样本转换成秩样本后，再对这个秩样本进行方差分析。

设有 k 组样本，n_i 是第 i 组样本中的观察数，n 是所有样本中的观察总数，R_{ij} 是第 i 组样本中的第 j 个观察值的秩值，$R_{i\bullet}$ 是第 i 组样本中的秩和。原假设为各组之间不存在差异，或者说各组的样本来自的总体具有相同的中心或均值或中位数。在原假设为真时，各组样本的秩平均应该与全体样本的秩平均 $\frac{1+2+\cdots+n}{n}=\frac{n+1}{2}$ 比较接近。因此组间平方和 $\mathrm{SSG}=\sum_{i=1}^{k}n_i\left(\frac{R_{i\bullet}}{n_i}-\frac{n+1}{2}\right)^2$，

同时全体样本秩方差为：

$$\begin{aligned}\sigma_R^2 &= \frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(R_{ij}-\frac{n+1}{2}\right)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}\left(i-\frac{n+1}{2}\right)^2 \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} i^2-\frac{n(n+1)^2}{4}\right) \\ &= \frac{1}{n-1}\left(\frac{n(n+1)(2n+1)}{6}-\frac{n(n+1)^2}{4}\right) \\ &= \frac{n(n+1)}{12}\end{aligned}$$

因此 Kruskal-Wallis 秩和统计量 KW 为：

$$\begin{aligned}\mathrm{KW} &= \frac{\mathrm{SSR}}{\sigma_R^2} \\ &= \frac{12}{n(n+1)}\sum_{i=1}^{k} n_i\left(\frac{R_{i\bullet}}{n_i}-\frac{n+1}{2}\right)^2 \\ &= \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{{R_{i\bullet}}^2}{n_i}-3(n+1)\end{aligned}$$

如果样本中存在打结情况，需要使用校正系数 C 调整 KW 统计量：

$$C = 1-\frac{\sum(\tau_j^3-\tau_j)}{n^3-n}$$

其中，τ_j 为第 j 个结值的个数。调整后的 KW_c 统计量为 $\mathrm{KW}_c = \mathrm{KW}/C$

若每组样本中至少有 5 个观察值，则样本统计量 KW_c 接近自由度为 $k-1$ 的卡方分布。因此卡方分布将用来决定 KW_c 统计量的显著性 P 值。

9.5.2 SAS 实例——探索不同专业学生英语成绩差异

例 9-4 在某次研究生公共英语考试后，分别随机抽取了 10 名工科、10 名理科和 10 名文科学生成绩，具体数据如表 9-12 所示，相应的 SAS 数据集在光盘中的存储路径为“data\chap9\score”。试分析不同专业的学生英语成绩是否存在显著的差异？（$\alpha = 0.05$）

表 9-12 不同专业公共英语成绩抽样表

工科（A）	112	54	132	110	109	105	122	87	74	93
理科（B）	89	95	112	115	123	117	84	85	92	47
文科（C）	113	142	134	133	125	110	90	86	83	75

编程法：

编写程序如下所示（其在光盘中的存储路径为“data\chap9\score”）：

```
ods graphics ;
proc npar1way data=chap9.score wilcoxon;
```

```
/*调用 npar1way 过程，并进行 wilcoxon 检验*/
var score;              /*定义分析变量为 score*/
class major;            /*定义分类变量为 major*/
run;
ods graphics off;
```

注释： Kruskal-Wallis 多样本位置检验为 Wilcoxon 两样本位置检验的推广，因此在 NPAR1WAY 过程中统一用 Wilcoxon 控制选项来指定这两种分析方法。

选择 Run|Submit 命令提交程序，以下分析主要输出结果。

表 9-13 为分组变量 major 的三个水平的得分情况，当 major 取 A 水平时，即对于工科生的成绩，对应的变量个数为 6，总得分为 146、平均得分为 12.6，在原假设（即 “不同专业的学生平均成绩没有差别”）下的期望得分为 155、期望标准差约为 22.72。其他水平的得分信息可类似得到。观察此表可知水平之间的得分差异较小。

表 9-13　分组变量 type 的 5 个水平的得分情况

Wilcoxon Scores (Rank Sums) for Variable score Classified by Variable major					
major	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
A	10	146.00	155.0	22.725245	14.600
B	10	139.50	155.0	22.725245	13.950
C	10	179.50	155.0	22.725245	17.950
Average scores were used for ties.					

表 9-14 为多变量的 Kruskal-Wallis 检验结果，检验的 λ^2 统计量为 1.1896，对应的自由度为 2，*P* 值为 0.5517，大于显著性水平 0.05，则接受原假设，认为不同专业的学生英语平均成绩没有显著的差异。

表 9-14　Kruskal-Wallis 检验结果

Kruskal-Wallis Test	
Chi-Square	1.1896
DF	2
Pr > Chi-Square	0.5517

图 9-3 为分组变量 major 不同水平之下的 Wilcoxon 得分图。观察可知三个水平的 Wilcoxon 得分差异不明显，此图进一步证实了 Kruskal-Wallis 检验结果。

本例也可以用 Analyst 菜单操作的方式得到检验结果，操作方式和例 9-3 完全一致，读者可参照设置，在此不赘述。

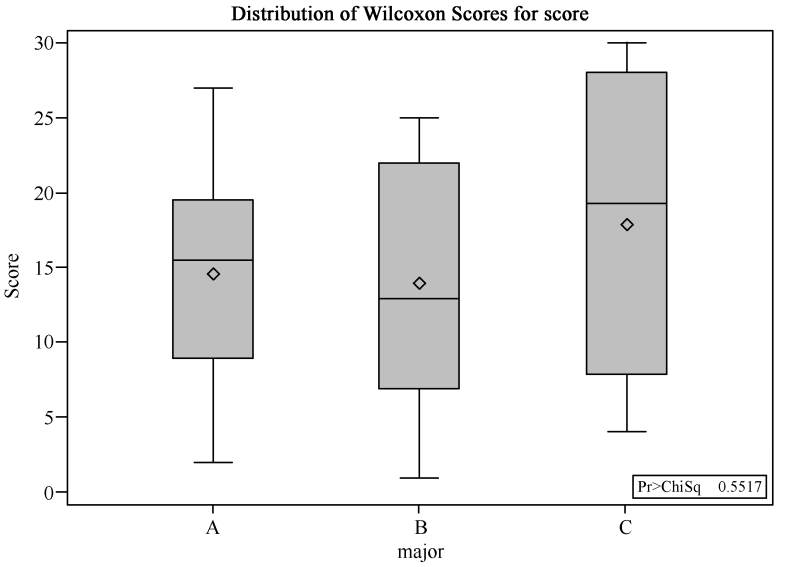


图 9-3 Wilcoxon 得分图

练习题

习题 9-1 某公司声称其减肥食谱在两个月以内可以减肥 6kg，20 个人使用这份食谱之后两个月所减少的重量如表 9-15 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap9\diet”）。请问两个月减 6kg 这种广告是否负责？

表 9-15 减重情况 单位：kg

6	1	6	5	7	2	5	6	4	5
3	6	3	8	0	8	6	2	4	4

（本习题的解答程序在光盘中的存储路径为“proc\chap9\diet”。）

习题 9-2 某制造商想要比较两种不同的生产方法所花费的生产时间（单位：min），随机地选取了 12 个工人，每一个工人都分别使用两种不同的生产方法来完成一项相同的任务，数据如表 9-16 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap9\worker”）。请问这两种生产方法所耗费的时间是否有显著的差异？

表 9-16 不同生产方法花费时间记录 单位：min

工人编号（ID）	方法一时间（Time1）	方法二时间（Time2）	工人编号（ID）	方法一时间（Time1）	方法二时间（Time2）
1	10.2	9.5	7	10.6	10.5
2	9.6	9.8	8	10.1	10.0
3	9.2	8.8	9	11.2	10.6

续表

工人编号（ID）	方法一时间（Time1）	方法二时间（Time2）	工人编号（ID）	方法一时间（Time1）	方法二时间（Time2）
4	10.6	10.1	10	10.7	10.2
5	9.9	10.3	11	10.6	9.8
6	10.2	9.3	12	10.6	10.4

（本习题的解答程序在光盘中的存储路径为“proc\chap9\worker”。）

习题 9-3　为了研究两个湖泊的环境对龟的生长的影响，释放了许多同样年龄的人工饲养的幼龟到两个湖中，每一只龟都带有记号。过了一段时间再打捞，在两个湖中发现有记号的龟的重量数据如表 9-17 所示（单位：kg），相应的 SAS 数据集在光盘中的存储路径为“data\chap9\tortoise”。请问两个湖泊的环境对做了记号的乌龟的重量增长是否有不同的影响？（α=0.05）

表 9-17　两个湖泊中乌龟增重情况　　单位：kg

湖泊 A	0.40	0.38	0.41	0.40	0.45	0.36	0.41	0.42	0.44
	0.47	0.39	0.42						
湖泊 B	0.47	0.48	0.42	0.44	0.41	0.45	0.46	0.50	0.44
	0.43								

（本习题的解答程序在光盘中的存储路径为“proc\chap9\tortoise”。）

习题 9-4　三个工厂的液晶电视显像管的寿命的具体的数据如表 9-18 所示（单位：月），相应的 SAS 数据集在光盘中的存储路径为“data\chap9\tube”。请分析不同工厂生产的液晶电视显像管的寿命是否有显著差异？（α=0.05）

表 9-18　不同工厂的液晶电视显像管的寿命　　单位：月

甲厂	52	49	48	52	56	53	55	54	
乙厂	51	47	49	51	53	48	49		
丙厂	53	49	59	58	57	54	55	53	56

（本习题的解答程序在光盘中的存储路径为“proc\chap9\tube”。）

第 10 章　主成分分析与因子分析

多元统计分析是统计学中内容丰富、应用性强的一个重要分支，它被广泛应用在自然科学、社会科学和经济学等各领域中，该分析主要包括主成分分析、因子分析、典型相关分析、判别分析和聚类分析。本章将介绍主成分分析和因子分析。

主成分分析（principal component analysis）——应用降维技术把多个变量化为少数几个主成分（即综合变量）的统计分析方法。

因子分析（factor analysis）——可看成主成分的推广和发展，是一种能较灵活地进行因子旋转的降维技术，主因子在降维后易于解释。

本章首先介绍实现主成分分析的 PRINCOMP 过程和实现因子分析的 FACTOR 过程，再依次介绍这两种分析方法的基本原理，并结合实例用编程和菜单实现分析。关于多元分析详细的原理介绍可参见《应用多元分析》（王学民编著，上海财经大学出版社）。

10.1　主成分分析

10.1.1　基本原理

主成分分析的目的是将多个变量化为少数几个相互独立的主成分。设有 n 组样品，每组样品有 p 个变量（如表 10-1 所示）。降维思想是利用 p 个变量来重新构造 q 个相互独立的综合变量（$q \leqslant p$），用较少的变量既尽可能地反映原来 p 个变量的统计特性。

表 10-1　p 个变量的 n 组样品数据

变量 / 样品	X_1	X_2	……	X_p
1	x_{11}	x_{12}	…	x_{1p}
2	x_{21}	x_{22}	…	x_{2p}
⋮	⋮	⋮	⋮	⋮
n	x_{n1}	x_{n2}	…	x_{np}

以下具体来求解主成分。假定 $x=(x_1,x_2,x_3,\cdots,x_p)'$ 为均值 $E(x)=\mu$、协方差矩阵 $D(x)=\boldsymbol{V}$ 的一组随机变量。考虑 $x_1,x_2,x_3,\cdots,x_p$ 的一个线性组合作为一个主成分：

$$Z=\boldsymbol{a}_1x_1+\boldsymbol{a}_2x_2+\cdots+\boldsymbol{a}_px_p=\boldsymbol{a}'x$$

其中，$\boldsymbol{a}'=(\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_p)$。为了让主成分尽可能多地保留原始变量的信息，选择在限制 $\boldsymbol{a}'\boldsymbol{a}=1$ 的条件下，寻找系数 $\boldsymbol{a}'=(\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_p)$ 使得 Z 的方差取最大值，即求 $\mathrm{Var}(\boldsymbol{a}'x)$ 的最大值。根据限制

性条件下的拉格朗日极值理论可以证明，在此情况下的 $\mathrm{Var}(\boldsymbol{a}'x)$ 的最大值等价于求 $\max\limits_{\boldsymbol{a}\neq 0}\dfrac{\boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{a}}$，就等于矩阵 $\boldsymbol{V}$ 的最大特征根 λ_1，$\boldsymbol{a}$ 就是 λ_1 对应的特征向量。记矩阵 $\boldsymbol{V}$ 的 p 个特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m > \lambda_{m+1} = \cdots = \lambda_p = 0$，且 m 个非零特征值所对应的特征向量分别为 $\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_m$，则

$$\max_{\boldsymbol{a}'\boldsymbol{a}=1} \boldsymbol{a}'\boldsymbol{V}\boldsymbol{a} = \lambda_1 = \boldsymbol{a}_1'\boldsymbol{V}\boldsymbol{a}_1$$

$$\max_{\substack{\boldsymbol{a}'\boldsymbol{a}=1\\ \boldsymbol{a}'\boldsymbol{a}_1=0}} \boldsymbol{a}'\boldsymbol{V}a = \lambda_2 = \boldsymbol{a}_2'\boldsymbol{V}\boldsymbol{a}_2$$

$$\vdots$$

$$\max_{\substack{\boldsymbol{a}'\boldsymbol{a}=1\\ \boldsymbol{a}'\boldsymbol{a}_i=0(i=1,2,\cdots,m-1)}} \boldsymbol{a}'\boldsymbol{V}\boldsymbol{a} = \lambda_m = \boldsymbol{a}_m'\boldsymbol{V}\boldsymbol{a}_m$$

那么把矩阵 $\boldsymbol{V}$ 的非 0 特征根 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m > 0$ 所对应的单位特征向量 $\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_m$ 分别作为 $x = (x_1, x_2, x_3, \cdots, x_p)'$ 的系数向量，$\boldsymbol{a}_1'x, \boldsymbol{a}_2'x, \cdots, \boldsymbol{a}_m'x$ 分别称为随机向量 x 的第 1 主成分、第 2 主成分，……，第 m 主成分。且当 $i \neq j$ 时 $\mathbf{Cov}(\boldsymbol{a}_i'x, \boldsymbol{a}_j'x) = \boldsymbol{a}_i'\boldsymbol{V}\boldsymbol{a}_j = \lambda_j\boldsymbol{a}_i'\boldsymbol{a}_j = 0$，即主成分之间是不相关的。在实际分析中，主要由观察数据阵 $\boldsymbol{X}$ 得到协方差 $\boldsymbol{V}$ 的估计 $\hat{\boldsymbol{V}}$，从 $\hat{\boldsymbol{V}}$ 出发计算它的特征值和特征向量，从而得到主成分。

$P_k = \lambda_k / \sum\limits_{j=1}^{P}\lambda_j$ 被定义为第 k 个主成分的贡献率，它反映了第 k 个主成分提取的全部信息量。$\sum\limits_{j=1}^{k}\lambda_j / \sum\limits_{j=1}^{p}\lambda_j$ 为前 k 个主成分的累积贡献率，它反映了前 k 个主成分共同提取的全部信息量。在实际分析中，如果前 m 个主成分的累积贡献率大于或等于 85%，则取 m 个主成分已经能够反映全部 p 个变量的绝大部分信息了。第 k 个主成分与 p 个变量 $x_1, x_2, \cdots, x_p$ 的系数矩阵称为因子载荷矩阵。

注意在实际分析中，变量的单位往往不一致，因此常将数据进行标准化处理，即使得第 i 个变量的均值为 0，方差为 1。设 $\overline{x}_i = \dfrac{1}{n}\sum\limits_{j=1}^{n}x_{ji}$，令

$$\tilde{x}_{ji} = \frac{x_{ji} - \overline{x}_i}{\sqrt{S_{ii}}} \quad i = 1, 2, \cdots, p;\ \ j = 1, 2, \cdots, n$$

称 $\tilde{x}_{ji}$ 为标准化后的数据。

10.1.2　SAS 过程——PRINCOMP 过程

PRINCOMP 过程的一般使用格式如下：

```
PROC  PRINCOMP  DATA=SAS 数据集 <选项列表>;
VAR 变量列表;
PARTIAL 变量列表;
FREQ 变量名;
WEIGHT 变量名;
BY 变量列表;
RUN;
```

注意：用 data=语句指定分析的资料类型可为原始数据集，也可为相关系数矩阵或协方差矩阵。

PROC PRINCOMP 语句后主要的控制选项如表 10-2 所示。

表 10-2　PROC PRINCOMP 语句后主要的控制选项

选　项	意　义
OUT=SAS 数据集	新建包括资料文件数据及主成分值的新输出数据集
OUTSTAT=SAS 数据集	新建包括以下统计量的输出数据集：MEAN（每一变量的均值）、STD（每一变量标准差）、N（观测个数）、CORR（相关系数矩阵）、COV（协方差矩阵）、EIGENVAL（特征根）、SCORE（特征向量）、SUMWGT（加权值的总和，仅当用户使用语句 PARTIAL，且指定 VARDEF=WDF 时输出）
NOINT	规定主成分不包括截距
COVARIANCE（COV）	指定以协方差\协方差矩阵作为分析的数据，若此项省略，则系统将以相关系数矩阵为依据
N=正整数	指定主成分个数
STANDARD（STD）	要求 OUT=数据集中含标准化的主成分，若此项省略，则系统输出未标准化的主成分
PREFIX=主成分名字	命名主成分
NOPRINT	抑制全部结果输出

PRINCOMP 过程中定义语句的意义如下：

VAR 语句——定义进行主成分分析的变量。若此语句缺失，程序中未被其他语句定义的所有数值变量均被纳入分析。

PARTIAL 语句——在计算其他变量相关系数或协方差矩阵时剔除由 PARTIAL 语句定义的变量的影响。

FREQ 语句——此语句定义的变量值代表资料文件内各观测重复数。

WEIGHT 语句——定义加权变量。

BY 语句——定义分组变量，使用前需预先将数据集按此语句定义的变量进行排序。

10.1.3　SAS 实例——客户信誉的“5C”评级分析

例 10-1　某企业为了了解客户的信用程度，评价客户的信用等级，采用常用的“5C”信用评级，以随时掌握客户违约的可能性，其中“5C”分别为：

A：品格，即客户的信誉。

B：资本，即客户的财务状况和实力。

C：即客户的偿还能力。

D：担保品，即客户的附加担保品价值。

E：环境，即客户的外部环境条件。

经过专家评定与打分，得到 20 个客户的 5 个项目的得分情况如表 10-3 所示，相应的 SAS 数据集在光盘中的存储位置为“data\chap10\custom_5C”，试对客户的信用等级进行评价。

表 10-3　客户“5C”信用评级得分表

ID	A	B	C	D	E	ID	A	B	C	D	E
1	76.4	82.5	73	72.8	72.5	11	85.3	76.2	89.3	86.4	78.5
2	68.6	73.2	67.3	65.2	75.5	12	94	94	87.5	89.5	90
3	89.7	89.3	93	80.5	85	13	80.6	68.9	70.8	68.8	63.4
4	77.3	73.4	69.9	67.8	73.8	14	56.7	58.4	53.4	60.8	65
5	85.2	68.3	68	60.2	74.5	15	70	69.2	71.7	63.9	68.9
6	83	81.2	80.2	84.3	75.5	16	76.5	82.5	78	77.8	71.7
7	93	92	82.5	88.5	82	17	70.6	75	69.6	65.1	72.5
8	83.6	67.9	78.8	63.8	65.4	18	92.7	85.3	92	89.5	85
9	88.7	89.3	92	83.5	86	19	79.5	70.6	68.9	73.8	70.8
10	79.5	74.6	77.9	78.8	75.8	20	84.6	68.3	72	63.2	79.5

编程法：

编写程序如下所示（其在光盘中的存储位置为“proc\chap10\custom_5C”）：

```
ods graphics on;
proc princomp data=chap10.Custom_5C(drop=ID);      /*调用 princomp 过程，且剔除变量 ID*/
 run;
ods graphics off;
```

选择 Run|Submit 命令提交程序，以下分析主要输出结果。

表 10-4 为 5 个变量的均值（Mean）和标准差（StD）。如变量 A 的均值为 80.775，标准差为 9.39450427，即被评分的 20 个客户在“人格信誉”这一项获得的平均评分为 80.775，且标准差约为 9.4。

表 10-4　变量均值和标准差

Simple Statistics					
	A	B	C	D	E
Mean	80.77500000	77.00500000	76.79000000	74.21000000	75.56500000
StD	9.39450427	9.59432261	10.31748133	10.33817148	7.35886683

表 10-5 为 5 个变量的相关系数矩阵，查阅此表可得到任意两个变量之间的相关系数，如变量 A（客户人格信誉评分）和变量 C（客户的财务状况和实力）之间的相关系数为 0.8267。

表 10-5　相关系数矩阵

Correlation Matrix					
	A	B	C	D	E
A	1.0000	0.7000	0.8267	0.6926	0.7201
B	0.7000	1.0000	0.8064	0.8440	0.8059
C	0.8267	0.8064	1.0000	0.8264	0.7532

续表

Correlation Matrix					
	A	B	C	D	E
D	0.6926	0.8440	0.8264	1.0000	0.7162
E	0.7201	0.8059	0.7532	0.7162	1.0000

表 10-6 为表 10-5 所示的相关矩阵的特征值：其中 Eigenvalue 列为特征值降序排列；Difference 列为两相邻特征值的差值，如此列的第一行对应的值 3.71763512（=4.07924345−0.36160833），依此类推计算；Proportion 列为对应特征值的贡献率，如第一个特征值的贡献率为 81.58%；Cumulative 为累计贡献率，如第二个特征值 0.36160833 对应的累计贡献率（即第一和第二个特征值的贡献率相加）为 88.82%。

表 10-6　相关矩阵的特征值

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.07924345	3.71763512	0.8158	0.8158
2	0.36160833	0.06280391	0.0723	0.8882
3	0.29880442	0.16372373	0.0598	0.9479
4	0.13508069	0.00981759	0.0270	0.9749
5	0.12526311		0.0251	1.0000

图 10-1 为主成分分析的碎石图和累计方差图，左图横轴代表主成分数（Principal Component）、纵轴代表特征值（Eigenvalue），观察发现从第二个特征值开始的变化幅度趋于平稳；右图横轴代表主成分数（Principal Component）、纵轴代表比率（Proportion），其中实线代表特征值的贡献率，虚线代表特征值的累计贡献率（Cumulative）。

结合表 10-6 所示结果，前两个主成分的累计贡献率达到了 88.82%，因此在本实验中取两个主成分即能够较好地解释原始数据的变异。

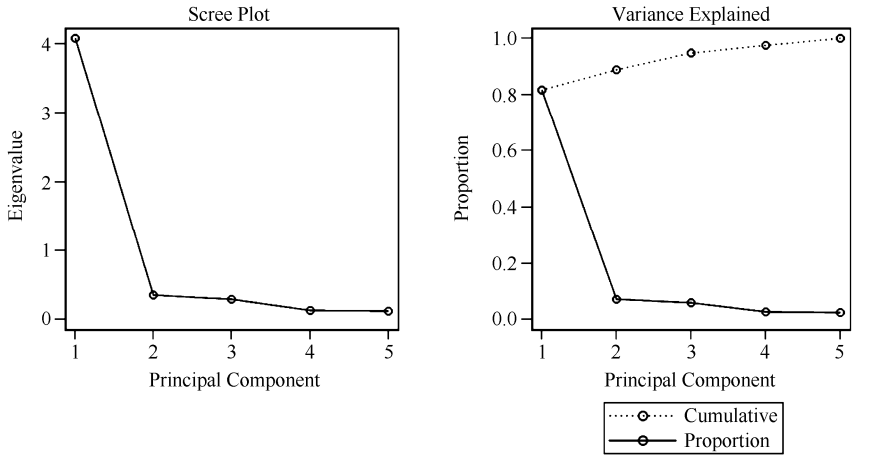


图 10-1　碎石图和累计方差图

表 10-7　主成分载荷矩阵

Eigenvectors					
	Prin1	Prin2	Prin3	Prin4	Prin5
A	0.430978	0.766689	−0.027176	0.358742	0.311469
B	0.456387	−0.441333	0.049320	0.692685	−0.338661
C	0.462472	0.214010	−0.304500	−0.467067	−0.655322
D	0.447928	−0.398540	−0.499843	−0.229142	0.581530
E	0.437550	−0.113047	0.808867	−0.347611	0.143776

表 10-7 为 5 个主成分的载荷矩阵。根据此表可写出第一主成分（Prin1）、第二主成分（Prin2）和 5 个变量的线性关系式：

Prin1=0.430978A+0.456387B+0.462472C+0.447928D+0.437550E

Prin2=0.766689A−0.441333B+0.214010C−0.398540D−0.113047E

观察发现第一主成分（Prin1）在各变量上面的载荷较为平均，都在 0.44 左右，则可将第一主成分解释为“客户的综合偿还能力”；而第二主成分（Prin2）在客户的内在能力（客户品格和偿还能力）上的载荷为正，而在客户的外在财务实力（资本、担保品价值和环境因素）方面上的载荷为负，则可将此主成分解释为“客户内在的素质”。

应用 ODS 图形输出系统还可绘制主成分之间的相关散点图，将本例的程序语句：

```
proc princomp data=chap10.Custom_5C(drop=ID);
```

改为：

```
proc princomp data=chap10.Custom_5C(drop=ID)plots=pattern(ncomp=2);
```

将绘制主成分 Prin1 和主成分 Prin2 之间的主成分与变量相关图，观察可知 5 个变量和第一主成分 Prin1 都呈现出正相关的关系，变量 A、C 和第二主成分 Prin2 呈较强正相关，而变量 E、D、B 和 Prin2 之间呈较强的负相关。

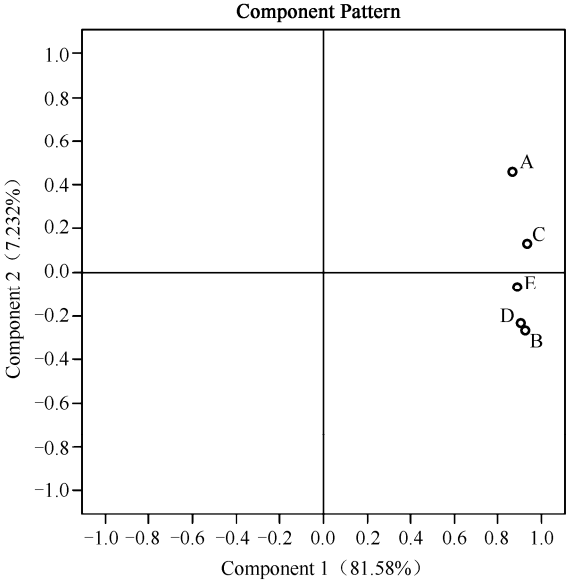


图 10-2　主成分图（Prin1 和 Prin2）

综上所述，本实验采用主成分分析的方法，根据原始变量的相关系数矩阵的前两个特征值的累积贡献率为 88.82%提取了两个主成分；由主成分和原始变量之间的线性关系得到主成分的实际意义解释：将第一主成分解释为“客户的综合偿还能力”，将第二主成分解释为“客户的内在素质”。最后给出了绘制主成分之间散点图的程序，读者可根据实际情况，酌情采用主成分图这一直观的分析结果表达方式。

菜单法：

步骤一：选择 Solutions|Analysis|Analyst 命令，进入 Analyst 分析主对话框。

步骤二：选择 File|Open By SAS Name|Chap10|custom_5C|OK 命令，打开数据集 chap10.custom_5C。

步骤三：选择 Statistics|Multivariate|Principal Components 命令，弹出如图 10-3 所示对话框，按住 Ctrl 键，单击选择变量 A 到 E 共 5 个变量，再单击 Variables（变量）按钮，将这些变量定义为主成分分析变量。

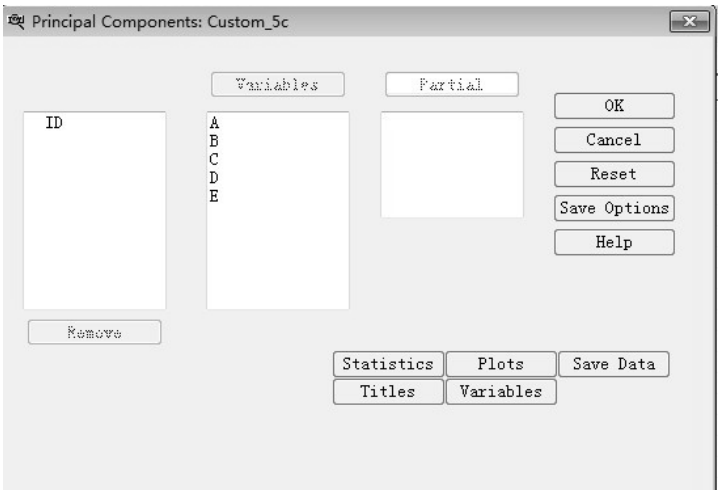


图 10-3　主成分分析对话框

步骤四：单击 Statistics（统计量）按钮，弹出如图 10-4 所示对话框，单击 Analyze（分析）选项框右侧的按钮，并单击下拉菜单中的 Correlations 选项（相关系数，此为系统默认设置）。在# of components（主成分个数）选项框内填入 5，设置保留 5 个主成分。单击 OK 按钮保存设置并返回如图 10-3 所示对话框。

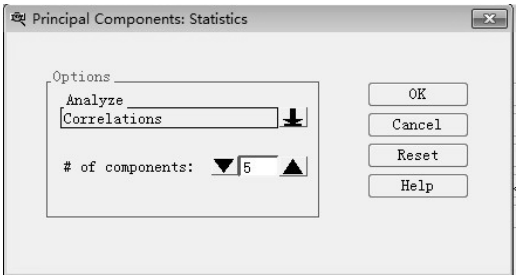


图 10-4　定义计算统计量

步骤五：单击 Plot（绘图）按钮，弹出如图 10-5 所示对话框，选择 Create scree plot（绘制

碎石图）选项，单击 For 选项框内的 Positive eigenvalues（实际特征值）选项，即设置绘制大于 1 的特征值的碎石图。若读者希望绘制所有特征值的碎石图可选择 All eigenvalues（所有特征值）。

单击 Component Plot（主成分图）标签，显示相应的选项卡，如图 10-6 所示。选择 Create component plots（绘制主成分图），用户可在 Dimensions（维度）选项框内自设主成分图的维度，系统默认绘制第一和第二主成分的主成分图，在 Id variable（识别变量）选项框内可设置观测的识别变量。本实验设置绘制第一和第二主成分的主成分图（如图 10-6 所示），单击 OK 按钮保存设置并返回如图 10-3 所示对话框。

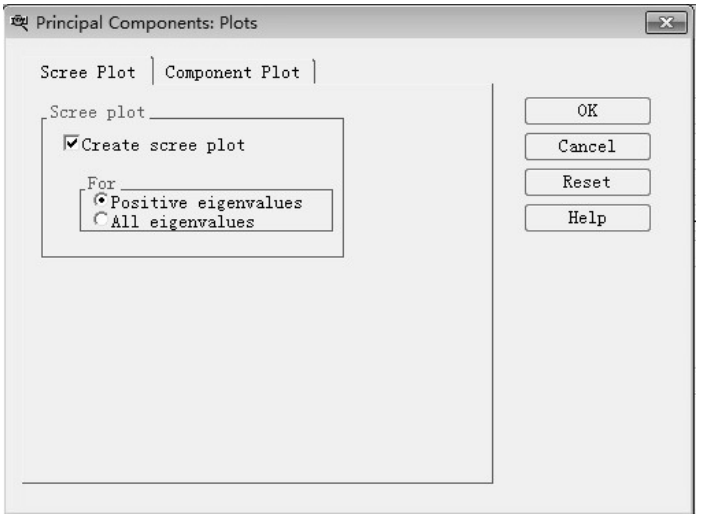


图 10-5　定义绘制碎石图

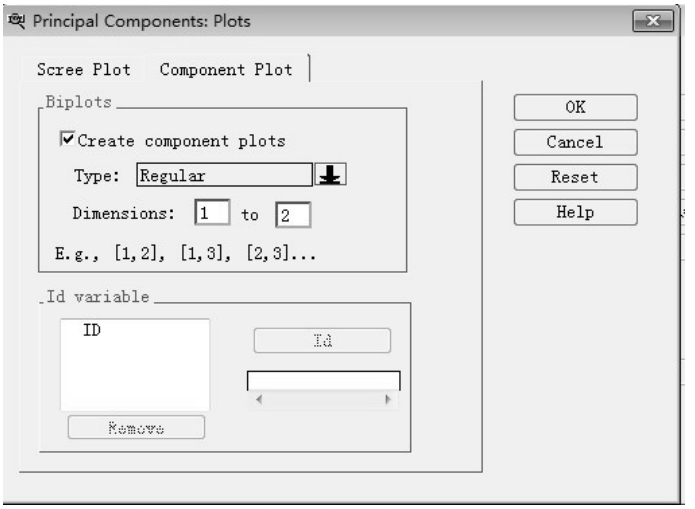


图 10-6　绘制主成分图

单击如图 10-3 所示对话框上的 OK 按钮，则系统将输出和编程方法同样的结果，请读者自行分析。

注意：在图 10-3 所示的对话框中单击 Save Data 按钮，可在弹出的对话框中进行保存（标准化）得分数据集和统计量数据集；单击 Titles 按钮，可在弹出的对话框中进行标题设置；单击

Variables 按钮，可在弹出的对话框中设置分层变量、加权变量和频数变量。读者可在实际应用中根据需要自行设置。

10.2 因子分析

10.2.1 基本原理

因子分析是主成分分析的推广，它也是把一些具有错综复杂关系的变量归结为少数几个综合因子的多变量统计分析方法。目的为用少量不可观察的潜在变量来解释原变量间的相关性或协方差关系。在这里我们把不可观察的潜在变量称为公共因子。假定对研究样品检测 p 个指标，p 个指标可能受到 m（$m < p$）个共同因素的影响，再加上其他对这些指标有影响的因素，用线性方程表示为：

$$\begin{cases} X_1 = a_{11}f_1 + a_{12}f_2 + \cdots + a_{1m}f_m + e_1 \\ X_2 = a_{21}f_1 + a_{22}f_2 + \cdots + a_{2m}f_m + e_2 \\ \cdots \\ X_p = a_{p1}f_1 + a_{p2}f_2 + \cdots + a_{pm}f_m + e_p \end{cases}$$

表示成矩阵形式为：$\underset{p\times 1}{\boldsymbol{X}} = \underset{P\times m}{\boldsymbol{A}}\ \underset{m\times 1}{\boldsymbol{f}} + \underset{p\times 1}{\boldsymbol{e}}$

f_i 被称为公共因子，$\boldsymbol{A}$ 被称为因子载荷矩阵，e_i 是变量 X_i 的特殊因子。设 $f_1, f_2, \cdots, f_m$ 分别是均值为 0、方差为 1 的随机变量，即 $\boldsymbol{D}(\boldsymbol{f}) = I_m$；特殊因子 $e_1, e_2, \cdots, e_p$ 分别是均值为 0、方差为 $d_1^2, d_2^2, \cdots, d_p^2$ 的随机变量，即 $\boldsymbol{D}(\boldsymbol{e}) = \mathrm{diag}(d_1^2, d_2^2, \cdots, d_p^2) = \boldsymbol{D}$；各特殊因子之间及特殊因子与公共因子相互独立，即 $\mathrm{Cov}(e_i, e_j) = 0, i \neq j$ 及 $\mathrm{Cov}(\boldsymbol{e}, \boldsymbol{f}) = 0$。$a_{ij}$ 是第 j 个变量在第 i 个公共因子上的负荷。

因子分析的目标是找出公共因子与特殊因子。在开始提取公共因子时，为简便起见还假定公共因子彼此不相关且具有单位方差。在这种情况下，向量 $\boldsymbol{X}$ 的协方差矩阵 $\boldsymbol{\Sigma}$ 可以表为 $\boldsymbol{\Sigma} = \boldsymbol{D}(\boldsymbol{X}) = \boldsymbol{D}(\boldsymbol{Af} + \boldsymbol{e}) = \boldsymbol{AA}' + \boldsymbol{D}$。

这里 $\boldsymbol{D}$=diag($d_1^2, d_2^2, \cdots, d_p^2$)，diag 表示对角矩阵。如果假定已经标准化，也就是说 $\boldsymbol{X}$ 的每一个分量 X_i 的均值都为 0，方差都是 1，即 $\boldsymbol{D}(X_i) = 1$，那么

$$\begin{cases} X_i = a_{i1}f_1 + a_{i2}f_2 + \cdots + a_{im}f_m + e_i \\ 1 = \mathrm{Var}(X_i) = \sum_{j=1}^{m} a_{ij}^2 + d_i^2 \end{cases}$$

记 $h_i^2 = \sum_{j=1}^{m} a_{ij}^2$，反映了公共因子 f 对 X_i 的影响，或者表述成反映了变量 X_i 对公共因子 f 的依赖程度，被称为公共因子 f 对 X_i 的“贡献”。

另外，还可以考虑指定的一个公共因子 f_j 对各个变量 X_i 的影响。实际上，f_j 对各个变量 X_i

的影响可由 $\boldsymbol{A}$ 中第 j 列的元素来描述，即 $g_j^2=\sum_{i=1}^{p}a_{ij}^2$，被称为公共因子 f_j 对 $\boldsymbol{X}$ 的“贡献”。显然 g_j^2 越大，f_j 对 $\boldsymbol{X}$ 的影响就越大，g_j^2 成为衡量因子重要性的一个尺度。

综上所述，矩阵 $\boldsymbol{A}$ 的统计意义如下：

- a_{ij} 是 X_i 和 f_j 的相关系数。
- $h_i^2=\sum_{j=1}^{m}a_{ij}^2$ 是 X_i 对公共因子 f 的依赖程度。
- $g_j^2=\sum_{i=1}^{p}a_{ij}^2$ 是公共因子 f_j 对 X 的各个分量总的影响。

以下介绍求解因子载荷矩阵 $\boldsymbol{A}$ 的方法。

如果已知 X 协方差矩阵 $\boldsymbol{\Sigma}$ 和 $\boldsymbol{D}$，可以很容易地求出 $\boldsymbol{A}$。

$$\boldsymbol{\Sigma}-\boldsymbol{D}=\boldsymbol{A}\boldsymbol{A}'$$

记 $\boldsymbol{\Sigma}^*=\boldsymbol{\Sigma}-\boldsymbol{D}$，则 $\boldsymbol{\Sigma}^*$ 是非负定矩阵。若记矩阵 $\boldsymbol{\Sigma}^*$ 的 p 个特征值 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m>\lambda_{m+1}=\cdots=\lambda_p=0$，且 m 个非零特征值所对应的特征向量分别为 $\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2,\cdots,\boldsymbol{\gamma}_m$，则 $\boldsymbol{\Sigma}^*$ 的谱分解式为

$$\begin{aligned}\boldsymbol{\Sigma}^*&=\lambda_1\boldsymbol{\gamma}_1\boldsymbol{\gamma}_1'+\lambda_2\boldsymbol{\gamma}_2\boldsymbol{\gamma}_2'+\cdots+\lambda_m\boldsymbol{\gamma}_m\boldsymbol{\gamma}_m'\\&=(\sqrt{\lambda_1}\boldsymbol{\gamma}_1,\sqrt{\lambda_2}\boldsymbol{\gamma}_2,\cdots,\sqrt{\lambda_m}\boldsymbol{\gamma}_m)(\sqrt{\lambda_1}\boldsymbol{\gamma}_1,\sqrt{\lambda_2}\boldsymbol{\gamma}_2,\cdots,\sqrt{\lambda_m}\boldsymbol{\gamma}_m)'\end{aligned}$$

则因子载荷矩阵 $\boldsymbol{A}=(\sqrt{\lambda_1}\boldsymbol{\gamma}_1,\sqrt{\lambda_2}\boldsymbol{\gamma}_2,\cdots,\sqrt{\lambda_m}\boldsymbol{\gamma}_m)$。

但在实际问题中，我们并不知道 $\boldsymbol{\Sigma}$、$\boldsymbol{D}$，即不知道 $\boldsymbol{\Sigma}^*$，仅知道有 n 个样品，每个样品测得 p 个指标。为了建立公因子模型，首先要估计因子载荷 a_{ij} 和特殊因子方差 d_i^2。常用的参数估计方法有以下三种：主成分法、主因子解法和极大似然法。

1）主成分法

主成分法求因子载荷矩阵 $\boldsymbol{A}$ 的具体方法如下：首先从资料矩阵出发求出样品的协方差矩阵，记为 $\hat{\boldsymbol{\Sigma}}$，其特征值为 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p\geqslant0$，相应单位正交特征向量为 $\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2,\cdots,\boldsymbol{\gamma}_p$，当后 $p-m$ 个特征值较小时，则对 $\hat{\boldsymbol{\Sigma}}$ 进行谱分解可以近似为

$$\hat{\boldsymbol{\Sigma}}=\lambda_1\boldsymbol{\gamma}_1\boldsymbol{\gamma}_1'+\lambda_2\boldsymbol{\gamma}_2\boldsymbol{\gamma}_2'+\cdots+\lambda_m\boldsymbol{\gamma}_m\boldsymbol{\gamma}_m'+\boldsymbol{D}$$

其中 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m>0$ 是协方差矩阵 $\hat{\boldsymbol{\Sigma}}$ 相应的前 m 个较大特征值。先取 $\boldsymbol{a}_1=\sqrt{\lambda_1}\boldsymbol{\gamma}_1$，然后看 $\hat{\boldsymbol{\Sigma}}-\boldsymbol{a}_1\boldsymbol{a}_1'$ 是否接近对角阵。如果接近对角阵，说明公共因子只要取一个就行了，所有指标主要受到这一个公共因子的影响；如果 $\hat{\boldsymbol{\Sigma}}-\boldsymbol{a}_1\boldsymbol{a}_1'$ 不是近似对角阵，再取 $\boldsymbol{a}_2=\sqrt{\lambda_2}\boldsymbol{\gamma}_2$，看 $\hat{\boldsymbol{\Sigma}}-\boldsymbol{a}_1\boldsymbol{a}_1'-\boldsymbol{a}_2\boldsymbol{a}_2'$ 是否接近对角阵，如果接近对角阵，就取两个公共因子；否则再取 $\boldsymbol{a}_3=\sqrt{\lambda_3}\boldsymbol{\gamma}_3$ 等，直到满足“要求”为止。这里的“要求”要视具体情况而定，一般而言，就像主成分分析一样，直接取前 q 个特征值和特征向量，使得它们的特征值之和占全部特征值之和的 85%以上即可。此时，特殊因子方差 $d_i^2=\hat{\Sigma}_{ii}-\sum_{t=1}^{q}a_{ti}^2(i=1,2,\cdots,p)$。

2）主因子解法

主因子解法是主成分法的一种修正，它是从资料矩阵出发求出样品的相关矩阵 $\boldsymbol{R}$，设 $\boldsymbol{R}=\boldsymbol{A}\boldsymbol{A}'+\boldsymbol{D}$，则 $\boldsymbol{R}-\boldsymbol{D}=\boldsymbol{A}\boldsymbol{A}'$。如果我们已知特殊因子方差的初始估计 $(\hat{d}_i^*)^2$，也就是已知了先验公因子方差的估计为 $(\hat{h}_i^*)^2=1-(\hat{d}_i^*)^2$，则约相关阵 $\boldsymbol{R}^*=\boldsymbol{R}-\boldsymbol{D}$ 为

$$\boldsymbol{R}^*=\begin{bmatrix}(\hat{h}_1^*)^2 & r_{12} & \cdots & r_{1p}\\ r_{21} & (\hat{h}_2^*)^2 & \cdots & r_{2p}\\ \vdots & \vdots & \vdots & \vdots\\ r_{p1} & r_{p2} & \cdots & (\hat{h}_p^*)^2\end{bmatrix}$$

计算 $\boldsymbol{R}^*$ 的特征值和特征向量，取前 m 个正特征值 $\lambda_1^*\geqslant\lambda_2^*\geqslant\cdots\geqslant\lambda_m^*\geqslant 0$ 及相应特征向量为 $\boldsymbol{\gamma}_1^*,\boldsymbol{\gamma}_2^*,\cdots,\boldsymbol{\gamma}_m^*$，则有近似分解式 $\boldsymbol{R}^*=\boldsymbol{A}\boldsymbol{A}'$，其中 $\boldsymbol{A}=(\sqrt{\lambda_1^*}\boldsymbol{\gamma}_1^*,\sqrt{\lambda_2^*}\boldsymbol{\gamma}_2^*,\cdots,\sqrt{\lambda_m^*}\boldsymbol{\gamma}_m^*)$，令 $\hat{d}_i^2=1-\sum_{t=1}^{m}a_{ii}^2(i=1,2,\cdots,p)$，则 $\boldsymbol{A}$ 和 $\boldsymbol{D}^*=\mathrm{diag}(\hat{d}_1^2,\hat{d}_2^2,\cdots,\hat{d}_p^2)$ 为因子模型的一个解，这个解就称为主因子解。

以上推导假设已知特殊因子方差的初始估计 $(\hat{d}_i^*)^2$，那么特殊因子方差的初始估计值如何得到呢？由于在实际中特殊因子方差 d_i^2（或公因子方差 h_i^2）是未知的，以上得到的解是近似解。为了得到更精确解，常常采用迭代主因子法，即利用上面得到的 $\boldsymbol{D}^*=\mathrm{diag}(\hat{d}_1^2,\hat{d}_2^2,\cdots,\hat{d}_p^2)$ 作为特殊方差的初始估计，重复上述步骤，直到解稳定为止。

公因子方差常用的初始估计有下面三种方法：

- h_i^2 取为第 i 个变量与其他所有变量的多重相关系数的平方（或者取 $d_i^2=1/r^{ii}$，其中 r^{ii} 是相关矩阵 $\boldsymbol{R}$ 的可逆矩阵 $\boldsymbol{R}^{-1}$ 的对角元素，则 $h_i^2=1-d_i^2$）。
- h_i^2 取为第 i 个变量与其他所有变量相关系数绝对值的最大值。
- 取 $h_i^2=1$，它等价于主成分解。

3）极大似然法

假定公共因子 f 和特殊因子 e 服从正态分布，那么我们可得到因子载荷矩阵和特殊方差的极大似然估计。设 p 维的 n 个观察向量 $\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)},\cdots,\boldsymbol{x}_{(n)}$ 为来自正态总体 $N_p(\mu,\Sigma)$ 的随机样本，则样本似然函数为 μ 和 $\boldsymbol{\Sigma}$ 的函数 $L(\mu,\boldsymbol{\Sigma})$。设 $\boldsymbol{\Sigma}=\boldsymbol{A}\boldsymbol{A}'+\boldsymbol{D}$，取 $\mu=\bar{x}$，对于一组确定的随机样本，μ 已经变成了已知的值，则似然函数 $L(\mu,\boldsymbol{\Sigma})$ 可以转换为 $\boldsymbol{A}$ 和 $\boldsymbol{D}$ 的函数 $\varphi(\boldsymbol{A},\boldsymbol{D})$。以下求使得函数 $\varphi(\boldsymbol{A},\boldsymbol{D})$ 能达到最大的 $\boldsymbol{A}$ 和 $\boldsymbol{D}$ 的值。为了保证得到唯一解，可以附加唯一性条件 $\boldsymbol{A}'\boldsymbol{D}^{-1}\boldsymbol{A}=$ 对角阵，再用迭代方法可求得极大似然估计的 $\boldsymbol{A}$ 和 $\boldsymbol{D}$ 的值。

因子模型被估计后，还必须对得到的公共因子 f 进行解释。即对每个公共因子给出一种意义明确的名称，这个公共因子的重要程度就是在因子模型矩阵中相应于这个因子的系数，显然这个因子的系数绝对值越大越重要，而接近 0 则表示对可观察变量没有什么影响。设 p 维可观察变量 $\boldsymbol{X}$ 满足因子模型 $\boldsymbol{X}=\boldsymbol{A}f+e$。设 $\boldsymbol{\Gamma}$ 是任一正交阵，则因子模型可改写为

$$\boldsymbol{X}=\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{\Gamma}'f+e\hat{=}\boldsymbol{A}^*f^*+e$$

其中，$\boldsymbol{A}^*=\boldsymbol{A}\boldsymbol{\Gamma}$，$f^*=\boldsymbol{\Gamma}'f$。

根据我们前面假定：每个公共因子的均值为 0，即 $E(f)=0$，每个公共因子的方差为 1，即 $D(f)=I$，各特殊因子之间及特殊因子与公共因子之间都是相互独立的，即 $\mathrm{Cov}(e_i,e_j)=0(i\neq j)$ 及 $\mathrm{Cov}(e,f)=0$。可以证明：

$$E(f^*)=E(\boldsymbol{\Gamma}'f)=\boldsymbol{\Gamma}'E(f)=0$$
$$D(f^*)=D(\boldsymbol{\Gamma}'f)=\boldsymbol{\Gamma}'D(f)\boldsymbol{\Gamma}=\boldsymbol{\Gamma}'I\boldsymbol{\Gamma}=I$$
$$\mathrm{Cov}(e,f^*)=\mathrm{Cov}(e,\boldsymbol{\Gamma}'f)=\boldsymbol{\Gamma}'\mathrm{Cov}(e,f)=0$$
$$D(\boldsymbol{X})=D(\boldsymbol{A}^*f^*+e)=D(\boldsymbol{A}^*f^*)+\mathrm{D}(e)=\boldsymbol{A}^*(\boldsymbol{A}^*)'+\boldsymbol{D}$$

因此 $X=\boldsymbol{AA}'+\boldsymbol{D}=\boldsymbol{A}^*(\boldsymbol{A}^*)'+\boldsymbol{D}$。说明如果 $\boldsymbol{A}$ 和 $\boldsymbol{D}$ 是一个因子解，任给正交阵 $\boldsymbol{\Gamma}$，$\boldsymbol{A}^*=\boldsymbol{A\Gamma}$ 和 $\boldsymbol{D}$ 也是因子解。由于正交阵 $\boldsymbol{\Gamma}$ 是任给的，所以因子解不是唯一的。在实际工作中，为了使载荷矩阵有更好的实际意义，在求出因子载荷矩阵 $\boldsymbol{A}$ 后，再右乘一个正交阵 $\boldsymbol{\Gamma}$，即变换了因子载荷矩阵，进行因子轴的正交旋转。

由于一个所有系数接近0或±1的旋转模型矩阵比系数多数为0与±1之间的模型容易解释，因此大多数旋转方法都是试图最优化模型矩阵的函数。在初始因子提取后，这些公因子是互不相关的。如果这些因子用正交变换（orthogonal transformation）进行旋转，旋转后的因子也是不相关的。如果因子用斜交变换（oblique transformation）进行旋转，则旋转后的因子变为相关的。但斜交旋转常常产生比正交旋转更有用的模型。

10.2.2　SAS 过程——FACTOR 过程

FACTOR 过程使用的一般格式如下：

```
PROC  FACTOR DATA=SAS 数据集 <选项列表>;
PRIORS 变量共通性的预估值;
VAR  变量列表;
PARTIAL  变量列表;
FREQ  变量名;
WEIGHT  变量名;
BY  变量列表;
RUN;
```

PROC FACTOR 语句后的控制选项被分成 5 类：定义数据集选项（如表 10-8 所示）、提取因子选项（如表 10-9 所示）、坐标转换选项（如表 10-10 所示）、控制报表输出选项（如表 10-11 所示）、其他选项（如表 10-12 所示）。

表 10-8　数据集定义选项

选　项	意　义
OUT=SAS 数据集	指定输出数据集，其中包括输入数据集的数据及因子得分（FACTOR SCORES），其值以 FACTOR1、FACTOR2 来表示。输入的数据集必须为原始数据集，不能为相关或协方差矩阵
OUTSTATA=SAS 数据集	指定包括包含样本均值、标准差、相关系数、协方差矩阵、特征根、因子模型等计算结果的 SAS 输出数据集

在 PROC FACTOR 语句过程中使用 METHOD=（或 M=）选项指定提取因子的方法，系统默认方法为 M=PRINCIPAL，若输入数据的格式为 TYPE=FACTOR 时，系统默认值为

METHOD=PATTERN。表 10-9 列举了供选的因子提取方法。

表 10-9　因子提取方法

选　　项	意　　义
M=PRINCIPAL（PRIN 或 P）	此因子提取方法视选项 PRIORS=而定。当此选项不与 PRIORS=并用，或者与 PRIORS=ONE 并用时，它的因子提取方法是主成分分析法；否则它的因子提取法为主轴因子分析法
M=PRINIT	指定迭代主因子分析法
M=ULS（或 U）	指定使用不加权的最小二乘因子分析法
M=ALPHA（或 A）	指定使用阿尔法因子分析
M=ML（或 M）	指定使用极大似然法进行因子分析，要求相关系数矩阵满秩
M=HARRIS（或 H）	指定采用 HARRIS 于 1962 年提出的主轴分析法，分析前提为相关系数矩阵为满秩
M=IMAGE（或 I）	针对映像共变异数矩阵做主成分分析
M=PATTERN	从输入数据文件（TYPE=FACTOR CORR 或 COV）内取得因子载荷矩阵。若因子间存在线性相关，则其间的相关系数也必须同时输入（TYPE='FCORR'的数据）
M=SCORE	从输入数据文件（TYPE=FACTOR CORR 或 COV）内取得因子分数的系数，这个输入数据集必须同时包括变量间的相关系数或其协方差矩阵

表 10-10　坐标转换选项

选　　项	意　　义
ROTATE（或 R）=坐标转换法	指定因子旋转方法，有如下方法供选择： （1）R=VARIMAX（或 V）：指定正交方差最大旋转，与 ROTART=ORTHOMAX 且 GAMMA=1 对应。 （2）R=QUARTIMAX（或 Q）：执行正交 4 次方最大旋转。 （3）R=BIQUARTIMAX（或 BIQMAX）：执行正交 8 次方最大旋转，与 ROTATE=ORTHOMAX 对应。 （4）R=EQUAMAX（或 E）：执行正交均方最大旋转，与 ROTATE=ORTHOMAX 且 GAMMA=因子个数对应。 （5）R=ORTHOMAX：执行标准正交转换法，其加权值来自选项 GAMMA= 。 （6）R=NONE（N）：不执行任何坐标转换，是 R=的系统默认值
PREROTATE（或 PRE）=坐标转换法	为 ROTATE=PROMAX 选项指定预先旋转的方法，此选项不可与 PROMAX 或 PROCRUSTES 联用，当 METHOD=PRTTERN 时，PREROTATE 必须是 NONE

表 10-11　控制报表输出选项

选　　项	意　　义
SIMPLE（或 S）	显示均值和标准差
CORR（或 C）	显示相关系数或偏相关系数矩阵
SCREE	将特征根从大到小排列后以图形显示，此图形称为 SCREE PLOT
PRINT	显示输入数据文件中的有关因子模型、得分系数及其他统计量。该选项只适合与 METHOD=PATTERN 或 SCORE 联用
RESIDUALS（或 RES）	显示残差相关矩阵及相应的偏相关矩阵，残差矩阵等于原始相关系数矩阵减去由因子形态导出的估计值矩阵

续表

选　项	意　义
PREPLOT	显示尚未经过坐标转换的因子载荷矩阵
PLOT	显示经过坐标转换后的因子载荷矩阵
NPLOT=正整数	显示前几个最重要的因子载荷矩阵。最小值为 2，默认值等于所有因子的总个数
SCORE	显示出因子得分系数
ALL	显示除 PLOT、NPLOT 以外其他选项所产生的报表
REORDER（或 RE）	让各种因子系数矩阵的这些行重新排列，使那些在第一因子上载荷量的绝对值高的变量排在前面几列，以协助解释因子的含义，输出数据集中的变量顺序不变

表 10-12　其他选项

选　项	意　义
NOINT	不使用截距项
NOCORR	与 METHOD=PATTERN 或 SCORE 联用，阻止相关系数被纳入 OUTSTAT=输出数据集内，当数据集含有许多变量但因子很少时，此选项可大大减少计算机的负荷
SINGULAR（或 SING）=正实数	指定一个矩阵不满秩的标准，默认值为 8～10
VARDEF=分母	为方差及协方差计算指定分母，可取值为 N、DF、WEIGHT、WGT、WDF，默认值为 DF

FACTOR 过程中一些语句的含义如下：

PRIORS 变量共通性的预估语句——此语句为每一个变量指定一个从 0.0～1.0 之间的初始共性方差估计值。估计值的数据应与 VAR 指令中变量的个数对应。例如：

```
PROC FACTOR;
VAR X Y Z;
PRIORS 0.5 0.6 0.7;
RUN;
```

该程序中 X 的共通性预估值为 0.5，Y 变量的共通性预估值为 0.6，Z 变量的共通性预估值为 0.7。

VAR 语句——定义参与因子分析的变量。若此语句缺失，则程序中未被其他语句定义的所有数值型变量均被纳入因子分析。

PARTIAL 语句——指定一组变量，其值将在其余变量中净化出来，由此得到值构成偏相关系数矩阵，而不是相关系数矩阵。

10.2.3　SAS 实例——我国各省市发展情况分析

例 10-2　从《中国统计年鉴 2010》中 2009 年的省会城市和计划单列市主要经济指标附表中选取了以下 10 个指标进行分析，具体数据如表 10-13 所示，相应的 SAS 数据集在光盘中的存储路径为“data\chap10\city”。试用恰当的综合指标来评价城市的发展状况。

X1：总人口，单位为万人；

X2：地区生产总值，单位为亿元；

X3：地方财政预算内收入，单位为万元；

X4：地方财政预算内支出，单位为万元；
X5：固定资产投资总额，单位为万元；
X6：在岗职工平均工资，单位为元；
X7：社会商品零售总额，单位为万元；
X8：普通高等学校在校学生数，单位为人；
X9：医院、卫生院个数；
X10：三废综合利用产品产值，单位为万元。

表 10-13　城市经济发展指标

city	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
北京	1246	12153	20268089	23193658	48584051	58140	53098869	577154	638	71680
天津	980	7522	8219916	11242778	50063247	45075	24308297	405968	437	187882
石家庄	977	3001	1259614	2409171	24363602	27370	11905536	366531	397	65087
太原	365	1545	1175322	1599051	7820157	33140	7216982	323321	258	74284
呼和浩特	227	1644	1067947	1651584	8348102	33993	6412127	203891	144	26489
沈阳	717	4269	3202070	4758822	35199470	38577	17785858	341863	301	20004
大连	585	4350	4002340	4711648	31136950	38765	13967483	236784	225	31868
长春	757	2849	1426506	3060062	22915489	30448	10893752	359423	309	77191
哈尔滨	992	3175	1933598	3484127	18920970	29251	15078539	468903	457	49001
上海	1400	15047	25402974	29896500	52733299	63549	51732408	512809	936	161409
南京	630	4230	4345080	4612662	26479817	43623	19354933	773394	208	196510
杭州	683	5088	5207899	4903983	22916543	43947	18049303	394087	324	1586099
宁波	571	4329	4328003	5060788	20042179	39138	14296750	135098	250	278885
合肥	491	2102	1808977	2458575	24684233	34144	7034168	352091	223	31899
福州	638	2604	1952612	2050925	16467177	30704	13386447	265682	216	33844
厦门	177	1737	2405608	2680527	8821159	36455	5661225	131451	49	76761
南昌	497	1838	1158798	1817495	14793151	30450	6344337	484890	186	31901
济南	603	3351	2101923	2599178	16553668	35661	15956509	632572	281	126297
青岛	763	4854	3770086	4335754	24588889	33257	17302231	269506	252	83646
郑州	731	3308	3019248	3530483	22890810	29837	14347614	617394	270	28949
长沙	652	3745	2463300	3140820	24417763	34888	15249091	502972	265	53816
广州	795	9138	7026527	7899155	26598516	49518	36157655	796006	253	111111
深圳	891	8201	8808168	10008394	17091514	46715	25679436	66952	101	38516
南宁	698	1525	1204628	2035519	10439120	32596	7570122	257576	202	69746
海口	158	490	384416	666116	2770332	30639	2771961	94153	73	1591
成都	1140	4503	3873626	6009694	40258902	34195	19499459	589291	562	171696
贵阳	367	972	1053636	1698423	7827910	27579	4127229	245768	219	36277
兰州	324	926	570385	1198342	5061800	28996	4697711	261847	159	119160

续表

city	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
西宁	194	501	281495	851191	3120429	28124	2015968	43782	105	10075
乌鲁木齐	241	1095	1135380	1338811	4120500	36500	4734172	130264	166	23025

编程法：

编写程序如下所示（其在光盘中的存储路径为“proc\chap10\city”）：

```
/*初步探索分析*/
ods graphics on;
proc factor data=chap10.city(drop=city)
                priors=smc
 /*设定取每一变量与其他变量的复相关系数平方值为该变量的共性方差的预估值*/
                plots=(scree);        /*绘制碎石图*/
run;
```

选择 Run|Submit 命令提交程序，以下分析主要结果输出。

表 10-14 为应用 SMC 方法对每个变量的共性方差的预估计值，如变量 X5 对应的共性方差预估值为 0.83727602。表 10-15 为约相关矩阵的特征值信息表：Eigenvalue 列为自上而下依次递减的特征值； Difference 列为两相邻特征值的差值；Proportion 列为特征值的贡献率；Cumulative 列为累计贡献率，前两个因子累积贡献率为 88.32%。表 10-15 的信息也可通过图 10-7 直观地显示出来。

表 10-14　共性方差先验估计

Prior Communality Estimates: SMC									
X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
0.91759701	0.98237967	0.99734398	0.99769567	0.83727602	0.94682570	0.98012314	0.67904873	0.92450165	0.59071650

表 10-15　约相关矩阵的特征值信息

Eigenvalues of the Reduced Correlation Matrix: Total = 8.85350806　Average = 0.88535081				
	Eigenvalue	Difference	Proportion	Cumulative
1	6.99466285	6.16997220	0.7900	0.7900
2	0.82469066	0.20928368	0.0931	0.8832
3	0.61540698	0.28580464	0.0695	0.9527
4	0.32960234	0.21440568	0.0372	0.9899
5	0.11519666	0.07004043	0.0130	1.0029
6	0.04515623	0.04742933	0.0051	1.0080
7	−0.00227310	0.01018986	−0.0003	1.0078
8	−0.01246296	0.00932169	−0.0014	1.0064
9	−0.02178465	0.01290231	−0.0025	1.0039
10	−0.03468696		−0.0039	1.0000

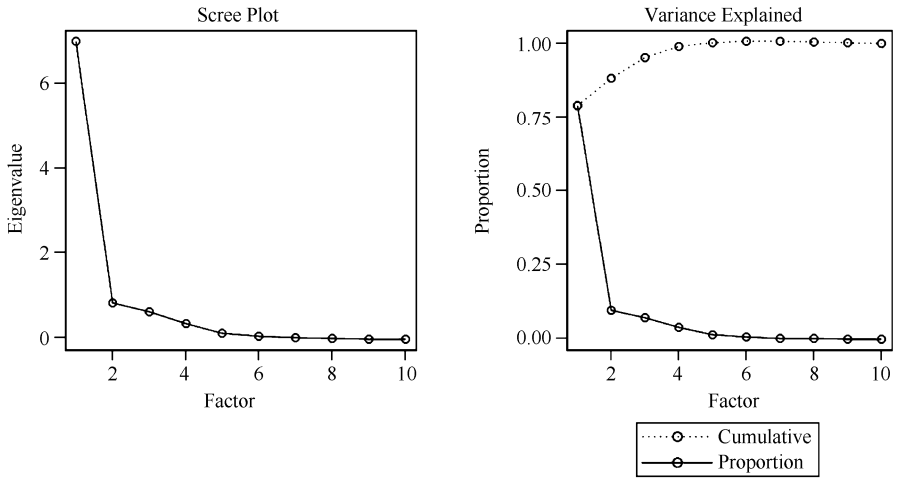


图 10-7　碎石图

因此本例确定选择两个主因子，修改程序如下所示：

```
/*精细分析*/
proc factor data=chap10.city(drop=city)
                priors=smc
/*设定取每一变量与其他变量的复相关系数平方值为该变量的共性方差的预估值*/
                rotate=promax reorder
/*进行最优斜交因子旋转，并将旋转后的因子按从大至小的顺序排列*/
                plots=(scree initloadings preloadings loadings);
/*绘制碎石图和原始因子旋转前、旋转后的因子载荷图*/
    nfactors=3;              /*规定取前两个因子*/
run;
```

选择 Run|Submit 命令提交程序，除了输出以上结果外，还将给出因子旋转以后的结果，以下分析主要结果。

表 10-16 为旋转之前的两个主因子（Factor1，Factor2）的变量 X1～X10 上的载荷信息。此信息也可由图 10-8 直观地显示出来。

表 10-16　因子载荷矩阵

Factor Pattern		
	Factor1	Factor2
X7	0.97463	−0.05053
X2	0.97271	−0.14544
X4	0.95484	−0.24400
X3	0.94792	−0.28859
X6	0.87546	−0.36049
X5	0.86888	0.24126
X1	0.86209	0.35420
X9	0.85864	0.27382

续表

Factor Pattern		
	Factor1	Factor2
X10	0.15721	−0.01742
X8	0.51164	0.51891

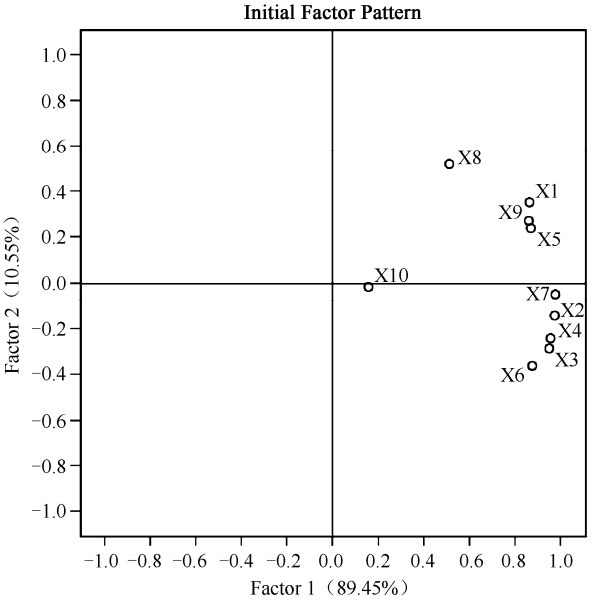


图 10-8　原始主因子载荷图

表 10-17 为主因子 Factor1 和 Factor2 解释的原始变量的方差。表 10-18 为最终共性方差的值。

表 10-17　主因子解释方差

Variance Explained by Each Factor	
Factor1	Factor2
6.9946629	0.8246907

表 10-18　共性方差值

Final Communality Estimates: Total = 8.434760									
X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
0.87327516	0.96736032	0.98694586	0.98853033	0.81342860	0.92354404	0.95256843	0.54682025	0.81620847	0.56607904

以上为因子旋转前提取出的两个主因子的信息，以下分别分析由正交转换法和最优斜交旋转以后的两个主因子信息。

表 10-19 为应用正交转换法的旋转矩阵。表 10-20 为旋转后两个因子 Factor1 和 Factor2 的因子载荷值，此表的信息也可通过图 10-9 显示出来。发现经过正交旋转后，两个主因子的载荷值都集中在了坐标的第二象限，则并不方便分别解释两个主因子。

表 10-19　正交转换矩阵

Orthogonal Transformation Matrix		
	1	2
1	0.81047	0.58578
2	−0.58578	0.81047

表 10-20　转换后因子载荷矩阵

Rotated Factor Pattern		
	Factor1	Factor2
X3	0.93731	0.32137
X6	0.92070	0.22065
X4	0.91680	0.36157
X2	0.87355	0.45192
X7	0.81951	0.52997
X10	0.13762	0.07797
X1	0.49121	0.79206
X9	0.53551	0.72489
X8	0.11071	0.72027
X5	0.56288	0.70450

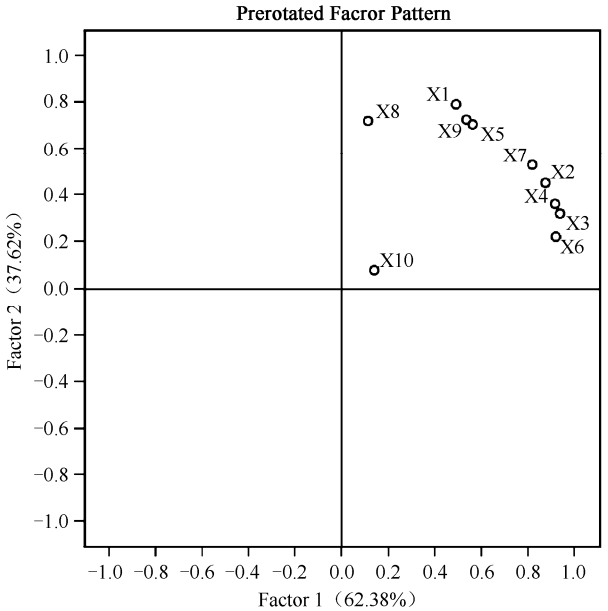


图 10-9　主因子图（正交旋转后）

表 10-21 为应用最优斜交旋转的旋转矩阵。表 10-22 为斜交旋转后 Factor1、Factor2 和 5 个变量之间的因子载荷。得到的因子相对正交旋转有较大的改进，主因子 1 在变量 X2、X3、X4、X6、X7 上载荷较大，而主因子 2 在变量 X1、X5、X8、X9 上载荷较大。此表的信息也可通过

图 10-10 反映出来。根据此表可得出主因子 1 的表达式：

Factor1=1.02998447X6+1.00490661X3+0.96160121X4+0.8677762X2+0.76625625X7
+0.13364939X10−0.1892083X8+0.24521243X1+0.32986366X9+0.37265197X5

表 10-21　斜交旋转矩阵

Procrustean Transformation Matrix		
	1	2
1	1.11242998	−0.4113193
2	−0.4091808	1.1363267

表 10-22　旋转后的因子载荷矩阵

Rotated Factor Pattern (Standardized Regression Coefficients)		
	Factor1	Factor2
X6	1.02998447	−0.1384242
X3	1.00490661	−0.0220127
X4	0.96160121	0.03651995
X2	0.8677762	0.16681375
X7	0.76625625	0.28679256
X10	0.13364939	0.0346078
X8	−0.1892083	0.83606661
X1	0.24521243	0.75501726
X9	0.32986366	0.65274793
X5	0.37265197	0.61550486

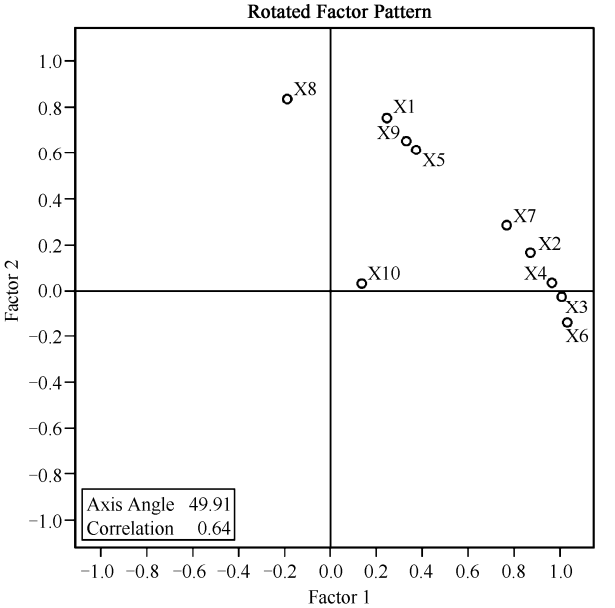


图 10-10　主因子图（斜交旋转后）

类似可得到主因子 2 的表达式。在本例中我们采用了主分量因子分析方法，得到了两个能高度概括原始数据信息的主因子，同时对因子进行了最大方差旋转和斜交旋转，发现斜交旋转都能得到较优的结果，最后可评定第一主因子在地区生产总值（X2）、地方财政预算内收入（X3）、地方财政预算内支出（X4）、在岗职工平均工资（X6）、社会商品零售总额（X7）上载荷较高，可将其概括为财政经济发展水平，第二主因子在总人口（X1）、固定资产投资总额（X5）、普通高校在校人数（X8）和医院、卫生院个数（X9）上载荷较高，可将其概括为城市发展规模因子。

菜单法：

步骤一：选择 Solutions|Analysis|Interactive Data Analysis 命令，进入 INSIGHT 模块界面（如图 10-11 所示）。单击 Library 选项框内的逻辑库 CHAP10，再单击数据集 CITY，单击 Open 按钮打开分析数据集 chap10.city。

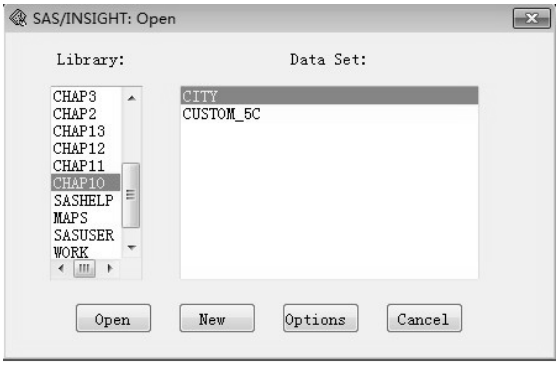


图 10-11　选择数据集

步骤二：选择 Analyze|Multivariate(Y X)命令，弹出如图 10-12 所示对话框，按住 Ctrl 键，拖动选择变量 X1～X10，再单击 Y 按钮，则将以上变量定义为分析变量。

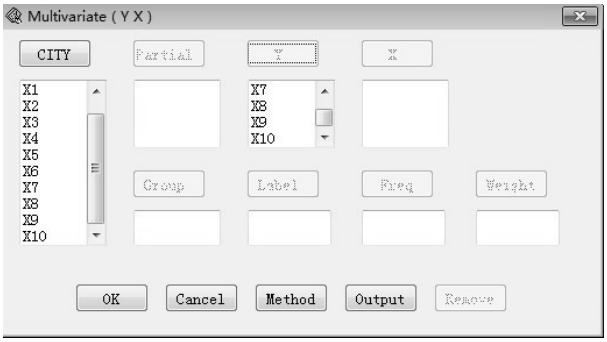


图 10-12　多元分析对话框

步骤三：单击 Output（输出）按钮，弹出如图 10-13 所示对话框，选择 Principal Component Analysis（主分量分析）复选框，再单击 Principal Component Options（主分量选项）按钮，弹出如图 10-14 所示对话框。选择 Component Rotation（因子旋转）复选框，再单击 Rotation Options（旋转选项）按钮，弹出如图 8-15 所示对话框。在此可定义旋转后的表格输出（Rotation Tables）和因子旋转图（Rotated Component Plots），本实验采用系统默认的输出方差最大化旋转矩阵（Orthogonal Rotation Matrix）和相关系数矩阵（Correlations），并单击选择输出标准化得分和原始变量的主因子图（Biplot(Std Y)）。依次单击对话框上的 OK 按钮保存设置最终返回如图 10-12

所示对话框。最后单击 OK 按钮，则将输出与编程法类似的结果，具体解释请参照以上分析。

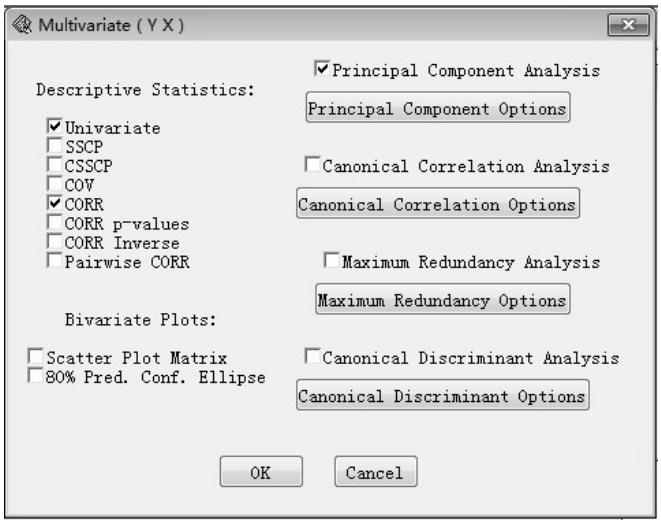


图 10-13　输出对话框

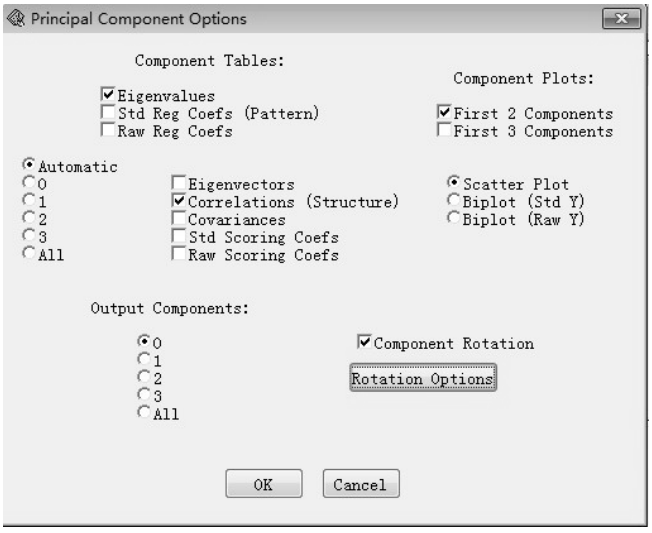


图 10-14　主分量选项

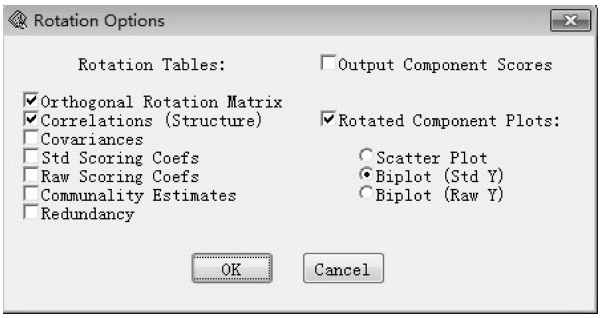


图 10-15　主分量旋转输出项

练习题

习题 10-1 已知我国 2009 年的各地区的主要财政支出，部分如表 10-23 所示，单位为亿元（包含完整数据的 SAS 数据集在光盘中的存储路径为“data\chap10\outcome”）。请用主成分分析方法得到这些支出主要包括哪些方面？

表中变量和对应的名目如下：

A1=国防
A2=公共安全
A3=教育
A4=科学技术
A5=文化教育与传媒
A6=社会保障和就业
A7=医疗卫生
A8=环境保护
A9=城乡社区事务
A10=农林水事务
A11=交通运输
A12=采掘电力信息等事务
A13=粮油物质储备等管理事务
A14=金融监管支出

表 10-23 各地区主要财政支出 单位：亿元

area	A1	A2	A3	A4	A5	A6	A7
北 京	5.29	161.38	365.67	126.31	74.75	234.29	166.63
天 津	1.02	73.49	173.61	34.00	19.81	115.90	54.22
河 北	4.51	151.20	439.33	26.43	38.02	317.42	174.68
area	A8	A9	A10	A11	A12	A13	A14
北 京	54.05	347.82	142.01	147.07	110.31	19.58	2.44
天 津	13.36	261.07	63.69	49.44	70.09	11.44	3.55
河 北	104.20	149.57	264.78	161.06	46.05	47.38	1.26

（本习题相应的解答程序在光盘中的存储路径为“proc\chap10\outcome”。）

习题 10-2 已知我国 1991—2009 年的居民消费价格指数（Index1），商品零售价格指数（Index2），工业品出厂价格指数（Index3），原材料、燃料、动力购进价格指数（Index4），固定资产投资价格指数（Index5），数据如表 10-24 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap10\index”）。请尝试用 1～2 个综合指标来概括这些指数。

表 10-24　年度价格指数

Year	Index1	Index2	Index3	Index4	Index5
1991	103.4	102.9	106.2	109.1	109.5
1992	106.4	105.4	106.8	111.0	115.3
1993	114.7	113.2	124.0	135.1	126.6
1994	124.1	121.7	119.5	118.2	110.4
1995	117.1	114.8	114.9	115.3	105.9
1996	108.3	106.1	102.9	103.9	104.0
1997	102.8	100.8	99.7	101.3	101.7
1998	99.2	97.4	95.9	95.8	99.8
1999	98.6	97.0	97.6	96.7	99.6
2000	100.4	98.5	102.8	105.1	101.1
2001	100.7	99.2	98.7	99.8	100.4
2002	99.2	98.7	97.8	97.7	100.2
2003	101.2	99.9	102.3	104.8	102.2
2004	103.9	102.8	106.1	111.4	105.6
2005	101.8	100.8	104.9	108.3	101.6
2006	101.5	101.0	103.0	106.0	101.5
2007	104.8	103.8	103.1	104.4	103.9
2008	105.9	105.9	106.9	110.5	108.9
2009	99.3	98.8	94.6	92.1	97.6

（本习题的解答程序在光盘中的存储路径为“proc\chap10\index”。）

第 11 章　典型相关分析

我们常用相关分析研究两个随机变量之间的相关关系，如计算 Pearson's 相关系数衡量连续变量的相关关系，计算 Spearman's 相关系数衡量等级变量的相关关系。当需要研究两组变量之间的相关关系时，则需要用到本章介绍的典型相关分析，这种多元统计分析方法能够有效揭示两组变量之间的相互线性依赖关系。在实际应用中，经常遇到考察学生的阅读速度、阅读理解能力和数学运算速度、解题能力的相关关系，学生文科类（历史、政治、语文、外语）和理科类（数学、物理、化学、生物）成绩的相关关系等问题都属于典型相关问题。

本章首先介绍典型相关分析的基本原理，然后介绍实现典型相关分析的 CONCORR 分析过程，最后介绍一个典型相关分析的实例。

11.1　基本原理

典型相关分析的基本思想是：为了研究两组随机变量 $(x_1,x_2,\cdots,x_p)$ 和 $(y_1,y_2,\cdots,y_q)$ 的相关关系，找出 $(x_1,x_2,\cdots,x_p)$ 的一个线性组合 u 及 $(y_1,y_2,\cdots,y_q)$ 的一个线性组合 v，使得 u 和 v 之间有最大可能的相关系数，以充分反映两组变量间的关系，以此将研究两组随机变量间相关关系转化为研究两个随机变量间的相关关系。如果一对变量 (u,v) 还不能完全表现出两组变量间的相关关系时，则继续寻找与已找出的变量对独立的变量对。

典型相关分析的数学模型如下：

设两组标准化的随机变量 $(x_1,x_2,\cdots,x_p)$ 和 $(y_1,y_2,\cdots,y_q)$，即 $E(x_i)=0,D(x_i)=1,\ i=1,2,\cdots,p$，$E(y_i)=0,D(y_i)=1,\ i=1,2,\cdots,q$，若记

$x=(x_1,x_2,\cdots,x_p)'$，$y=(y_1,y_2,\cdots,y_q)'$，此时 x 和 y 的协方差矩阵为：

$$D(x,y)=\begin{pmatrix}\boldsymbol{R}_{xx} & \boldsymbol{R}_{yx}\\ \boldsymbol{R}_{xy} & \boldsymbol{R}_{yy}\end{pmatrix}$$

其中，$\boldsymbol{R}_{xx}=D(x),\boldsymbol{R}_{yy}=D(y),\boldsymbol{R}_{xy}=\boldsymbol{R}_{yx}=\mathrm{Cov}(x,y)$

以下寻找 x 的线性组合 $u_1=\boldsymbol{l}_1'x$ 和 y 的线性组合 $v_1=\boldsymbol{m}_1'y$ 使 u_1 和 v_1 的相关系数 $\rho(u_1,v_1)$ 达到最大。由于对任意常数 a,b,c,d 有 $\rho(au_1+b,cv_1+d)=\rho(u_1,v_1)$（其中 $a\neq 0$，$c\neq 0$），因而假定 $D(u_1)=\boldsymbol{l}_1'\boldsymbol{R}_{xx}\boldsymbol{l}_1=1$，$D(v_1)=\boldsymbol{m}_1'\boldsymbol{R}_{yy}\boldsymbol{m}_1=1$，此时 $\rho(u_1,v_1)=\mathrm{Cov}(u_1,v_1)=\boldsymbol{l}_1'\boldsymbol{R}_{xy}\boldsymbol{m}_1$。在 $\boldsymbol{l}_1'\boldsymbol{R}_{xx}\boldsymbol{l}_1=1$ 与 $\boldsymbol{m}_1'\boldsymbol{R}_{yy}\boldsymbol{m}_1=1$ 条件下，使 $\boldsymbol{l}_1'\boldsymbol{R}_{xy}\boldsymbol{m}_1$ 达到最大的 $\boldsymbol{l}_1'$ 与 $\boldsymbol{m}_1'$ 分别与 x 和 y 组成的新变量

$$\begin{cases}u_1=\boldsymbol{l}_1'x\\ v_1=\boldsymbol{m}_1'y\end{cases}$$

称为第一对典型变量，其相关系数 $\rho(u_1,v_1)=\boldsymbol{l}_1'\boldsymbol{R}_{xy}\boldsymbol{m}_1$ 称为第一典型相关系数。若用一对变量还不

足以反映两组变量的相关性时，可继续定义第二对典型变量 $u_2=\boldsymbol{l}_2'x, v_2=\boldsymbol{m}_2'y$，这时除要求 $D(u_2)=1$，$D(v_2)=1$ 外，还要求第二对典型变量和第一对典型变量之间为独立的，即在条件 $\mathrm{Cov}(u_1,u_2)=0$，$\mathrm{Cov}(u_1,v_2)=0$，$\mathrm{Cov}(v_1,u_2)=0$ 和 $\mathrm{Cov}(v_1,v_2)=0$ 下使 $\rho(u_2,v_2)=\mathrm{Cov}(u_2,v_2)=\boldsymbol{l}_2'\boldsymbol{R}_{xy}\boldsymbol{m}_2$ 达到最大。推广到一般情形，称 $u_j=\boldsymbol{l}_j'x, v_j=\boldsymbol{m}_j'y$ 为第 j 对典型变量，其系数向量 $\boldsymbol{l}_j'$ 与 $\boldsymbol{m}_j'$ 使 $\boldsymbol{l}_j'\boldsymbol{R}_{xy}\boldsymbol{m}_j$ 达到最大，并且满足如下条件：

$$\begin{cases}\boldsymbol{l}_j'\boldsymbol{R}_{xx}\boldsymbol{l}_j=1\\ \boldsymbol{m}_j'\boldsymbol{R}_{yy}\boldsymbol{m}_j=1\\ \boldsymbol{l}_j'\boldsymbol{R}_{xx}\boldsymbol{l}_i=\boldsymbol{l}_j'\boldsymbol{R}_{xy}\boldsymbol{m}_i=\boldsymbol{m}_j'\boldsymbol{R}_{yx}\boldsymbol{l}_i=\boldsymbol{m}_j'\boldsymbol{R}_{yy}\boldsymbol{m}_i=0\end{cases}$$

$i=1,2,\cdots,j-1$，此时称 $\boldsymbol{l}_j'\boldsymbol{R}_{xy}\boldsymbol{m}_j$ 为第 j 对典型相关系数。

一般采用拉格朗日乘法，从 $j=1$ 开始逐一求 $\boldsymbol{l}_j$、$\boldsymbol{m}_j$。如求 $\boldsymbol{l}_1$、$\boldsymbol{m}_1$ 时首先假定 $\boldsymbol{R}$ 是正定矩阵，记

$$\phi(\boldsymbol{l}_1,\boldsymbol{m}_1)=\boldsymbol{l}_1'\boldsymbol{R}_{xy}\boldsymbol{m}_1-\frac{\lambda}{2}(\boldsymbol{l}_1'\boldsymbol{R}_{xx}\boldsymbol{l}_1-1)-\frac{\mu}{2}(\boldsymbol{m}_1'\boldsymbol{R}_{yy}\boldsymbol{m}_1-1)$$

其中 λ、μ 为 Lagrage 乘子，为使计算式用 $-\frac{\lambda}{2}$、$-\frac{\mu}{2}$ 表示。将 φ 对 $\boldsymbol{l}_1$、$\boldsymbol{m}_1$ 分别求偏导，并令其为 0，再与约束条件联立，则 $\boldsymbol{l}_1$、$\boldsymbol{m}_1$ 应满足以下方程组：

$$\begin{cases}\boldsymbol{R}_{xy}\boldsymbol{m}_1-\lambda\boldsymbol{R}_{xx}\boldsymbol{l}_1=0 & (1)\\ \boldsymbol{R}_{yx}\boldsymbol{l}_1-\mu\boldsymbol{R}_{yy}\boldsymbol{m}_1=0 & (2)\\ \boldsymbol{l}_1'\boldsymbol{R}_{xx}\boldsymbol{l}_1=1 & (3)\\ \boldsymbol{m}_1'\boldsymbol{R}_{yy}\boldsymbol{m}_1=1 & (4)\end{cases}$$

将方程组（1）式左乘 $\boldsymbol{l}_1'$ 得到 $\boldsymbol{l}_1'\boldsymbol{R}_{xy}\boldsymbol{m}_1-\lambda\boldsymbol{l}_1'\boldsymbol{R}_{XX}\boldsymbol{l}_1=0$ 又因为方程组（3）式 $\boldsymbol{l}_1'\boldsymbol{R}_{xx}\boldsymbol{l}_1=1$，则有 $\boldsymbol{l}_1'\boldsymbol{R}_{xy}\boldsymbol{m}_1=\lambda$，类似的将方程组（2）式左乘 $\boldsymbol{m}_1'$，并利用方程组（4）式得到 $\boldsymbol{m}_1'\boldsymbol{R}_{yx}\boldsymbol{l}_1=\mu$，由于 $\boldsymbol{R}_{xy}=\boldsymbol{R}_{yx}$，则 $\lambda=\mu$。将方程组中的 μ 用 λ 替换得到 $\boldsymbol{m}_1=\frac{1}{\lambda}\boldsymbol{R}_{yy}^{-1}\boldsymbol{R}_{yx}\boldsymbol{l}_1$，将其代入方程组（2）式，得到 $\boldsymbol{R}_{xy}\boldsymbol{R}_{yy}^{-1}\boldsymbol{R}_{yx}\boldsymbol{l}_1=\lambda^2\boldsymbol{R}_{xx}\boldsymbol{l}_1$，再由 $\boldsymbol{R}_{xx}$ 的非退化性知：

$$\boldsymbol{R}_{xx}^{-1}\boldsymbol{R}_{xy}\boldsymbol{R}_{yy}^{-1}\boldsymbol{R}_{yx}\boldsymbol{l}_1=\lambda^2\boldsymbol{l}_1$$

记 $\boldsymbol{M}_1=\boldsymbol{R}_{xx}^{-1}\boldsymbol{R}_{xy}\boldsymbol{R}_{yy}^{-1}\boldsymbol{R}_{yx}$，则 λ^2 是 $\boldsymbol{R}_{xx}^{-1}\boldsymbol{R}_{xy}\boldsymbol{R}_{yy}^{-1}\boldsymbol{R}_{yx}$ 的特征根，$\boldsymbol{l}_1$ 是其对应的特征向量。又知 $\lambda=\boldsymbol{l}_1'\boldsymbol{R}_{xy}\boldsymbol{m}_1$ 是 u_1 与 v_1 的相关系数，要求其达到最大，λ^2 一定是 $\boldsymbol{R}_{xx}^{-1}\boldsymbol{R}_{xy}\boldsymbol{R}_{yy}^{-1}\boldsymbol{R}_{yx}$ 的最大特征根，$\boldsymbol{l}_1$ 是最大特征根 λ^2 对应的特征向量；进而 $\boldsymbol{m}_1$ 可由关系式 $\boldsymbol{m}_1=\frac{1}{\lambda}\boldsymbol{R}_{yy}^{-1}\boldsymbol{R}_{yx}\boldsymbol{l}_1$ 求出。第一典型相关系数 λ_1 即为 $\boldsymbol{R}_{xx}^{-1}\boldsymbol{R}_{xy}\boldsymbol{R}_{yy}^{-1}\boldsymbol{R}_{yx}$ 的最大特征根的算术根。

类似的可以证明 $\boldsymbol{m}_1$ 是 $\boldsymbol{M}_2=\boldsymbol{R}_{yy}^{-1}\boldsymbol{R}_{yx}\boldsymbol{R}_{xx}^{-1}\boldsymbol{R}_{xy}$ 的最大特征根对应的特征向量。由于 $\boldsymbol{M}_1$ 与 $\boldsymbol{M}_2$ 有相同的非零特征根，因此此时求出的 $\boldsymbol{m}_1$ 和应用关系式 $\boldsymbol{m}_1=\frac{1}{\lambda}\boldsymbol{R}_{yy}^{-1}\boldsymbol{R}_{yx}\boldsymbol{l}_1$ 求出的 $\boldsymbol{m}_1$ 是一致的。

类似，可知 $\boldsymbol{l}_2$ 是 $\boldsymbol{M}_1$ 的第二大的特征根 λ_2^2 对应的特征向量，$\boldsymbol{m}_2$ 可通过 $\boldsymbol{m}_2=\frac{1}{\lambda_2}\boldsymbol{R}_{yy}^{-1}\boldsymbol{R}_{yx}\boldsymbol{l}_2$ 求出。

推广到一般情形，求出 $\boldsymbol{M}_1$ 的 r 个非零特征根 $\lambda_1^2\geqslant\lambda_2^2\geqslant\cdots\geqslant\lambda_r^2$，$\boldsymbol{M}_1$ 对应于这些特征根的特

征向量分别记为$\boldsymbol{l}_1,\boldsymbol{l}_2,\cdots,\boldsymbol{l}_r$，进而$\boldsymbol{m}_j=\frac{1}{\lambda_j}\boldsymbol{R}_{yy}^{-1}\boldsymbol{R}_{yx}\boldsymbol{l}_j$（$j=1,2,\cdots,r$），以$\boldsymbol{l}_j$、$\boldsymbol{m}_j$为系数可组成第 j 对典型变量$u_j=\boldsymbol{l}_j'x$，$v_j=\boldsymbol{m}_j'x$。第 j 对典型变量对应的相关系数λ_j是λ_j^2的算术根，这便是第 j 个典型相关系数，$j=1,2,\cdots,r$，这里$r\leqslant\min(p,q)$。

在实际应用中 $\boldsymbol{R}$ 通常是未知的，已知的只是$\begin{pmatrix}x\\y\end{pmatrix}$的 n 个样本，则需要根据样本估计 $\hat{\boldsymbol{R}}_{xx}$、$\hat{\boldsymbol{R}}_{xy}$和 $\hat{\boldsymbol{R}}_{yy}$。然后用相应的估计量代替 $\boldsymbol{R}$ 中相应的未知参数矩阵，计算过程与之前类似。

注意：到用这种方式求出的第一对典型相关包含最多的有关两组变量间相关的信息，第二对其次，其他对依次递减，各对典型相关所含的信息互不重复。而选取典型相关对数可通过检验“该典型相关系数为零”来确定。

11.2 SAS 过程——CANCORR 过程

在 SAS 系统中，用 CANCORR 过程进行典型相关分析，其一般使用格式如下：

```
PROC CANCORR  <选项列表>;
VAR 变量列表;
WITH 变量列表;
PARTIAL 变量列表;
FREQ 变量名;
WEIGHT 变量名;
BY 变量列表;
RUN;
```

PROC CANCORR 语句、VAR 语句和 WITH 语句是该过程必不可缺的，其他语句视情况选择。

PROC CANCORR 语句后的控制选项可分为以下 5 类：与数据集的输入、输出有关的选项（如表 11-1 所示）、控制报表打印的选项（如表 11-2 所示）、指定输出数据集名及计算过程的选项（如表 11-3 所示）、与回归分析有关的选项（如表 11-4 所示）、与回归分析中产生的统计量有关的选项（如表 11-5 所示）。

表 11-1　数据集输入、输出控制选项

选　项	意　义
DATA=SAS 数据集	指定输入数据集，此数据集可为 SAS 的原始数据集或为 TYPE=CORR、UCORR、COV、UCOV、SSCP 或 TYPE=FACTOR 的数据集
OUT=SAS 数据集	生成包含原始数据集和典型变量得分的输出 SAS 数据集，当输入的数据集的类型 TYPE=CORR、COV、FACTOR、SSCP、UCORR、UCOV 时，不可使用此选项
OUTSTAT=SAS 数据集	生成包含各种统计量的 SAS 数据集

表 11-2　控制报表输出的选项

选　项	意　义	选　项	意　义
SIMPLE（或 S）	输出简单描述性统计量	ALL	输出所有统计量
CORR（C）	输出原始变量之间的相关系数矩阵	NOPRINT	抑制所有输出
SHORT	输出典型相关系数与其 *F* 检验的 *P* 值		

表 11-3　控制输出数据集名及计算过程的选项

选　项	意　义
EDF=正整数	若输入数据集是某一个回归分析的结果，则指定 *F* 检验中的分母的自由度，默认值为有效观测值−1
RDF=正整数	若输入数据集是某一个回归分析的结果，则指定 *F* 检验中的分子的自由度，默认值为原回归分析中自变量的个数
NOINT	规定在典型相关分析与回归分析的模型中不包括截距，通常在截距项在模型中不显著的情况下使用
VPREFIX（或 VP）=典型变量的名字	为 VAR 语句指定的典型变量指定前缀名。若指定 VP=AB，则第一、第二典型变量就是 AB1、AB2，默认值为 V1、V2…
WPREFIX（或 WP）=典型变量的名字	为 WITH 语句指定的典型变量指定前缀名。默认值为 W1、W2…
VNAME（或 VN）='VAR 变量名称'	为 VAR 语句中所列出的变量名命名，名字字符限定在 40 个字符，需用英文输入状态下的单引号引出来

表 11-4　与回归有关的选项

选　项	意　义
VDEP	分别以 VAR 和 WITH 语句指定的变量为因变量和自变量进行多元回归分析
VREG	分别以 WITH 和 VAR 语句指定的变量为因变量和自变量进行多元回归分析

表 11-5　与回归统计量有关的选项

选　项	意　义
ALL	输出所有的统计量，并将其保存在 OUTSTAT=输出数据集中
INT	要求回归模型中包括截距项，此选项应与 B、SEB、T、PROBT 同时联用
B	要求打印回归系数
SEB	计算且输出回归系数的标准误
T	要求输出回归系数的 *t* 检验
PROBT	输出系数 *t* 检验对应的 *P* 值
STB	计算且输出标准化的回归系数
SMC	输出回归分析中产生的复相关系数的平方与 *F* 检验的结果
CORRB	输出回归系数估计值的相关系数矩阵
PCORR	输出自变量和因变量之间的偏相关系数
SPCORR	输出因变量和自变量之间的偏相关系数
SQPCORR	输出自变量和因变量之间的偏相关系数的平方
SQSPCORR	输出因变量和自变量之间的偏相关系数的平方

CONCORR 过程中使用的语句含义如下：

VAR 语句——指定进行典型相关分析的第一组数值型变量。若省略此语句，则未被其他语句定义的所有数值型变量将构成第一组变量。

WITH 语句——列举典型相关分析的两组变量之间的第二组数值型变量，此语句不可省略。

PARTIAL 语句——定义在偏相关的基础上进行典型相关分析，分析时首先固定 PARTIAL 语句中所指定变量的作用，然后对 VAR 和 WITH 语句中指定的两组变量进行偏相关分析。

WEIGHT 语句——指定观测的加权变量，此语句定义的变量必须为正值。

11.3 SAS 实例——生理指标和训练指标的相关分析

例 11-1 某健身俱乐部记录了 20 位中年男性的三项生理指标和三项训练指标，数据如表 11-6 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap11\fit”）。试分析健身者的身体素质和运动技能的相关性。（显著性水平 $\alpha = 0.1$ ）

表 11-6 生理指标和训练指标数据

生理指标			训练指标		
体重（Weight）	胸围（Waist）	心跳（Pulse）	拉链条个数（Chins）	仰卧起坐（Situps）	跳高（Jumps）
191	36	50	5	162	60
189	37	52	2	110	60
193	38	58	12	101	101
162	35	62	12	105	37
189	35	46	13	155	58
182	36	56	4	101	42
211	38	56	8	101	38
167	34	60	6	125	40
176	31	74	15	200	40
154	33	56	17	251	250
169	34	50	17	120	38
166	33	52	13	210	115
154	34	64	14	215	105
247	46	50	1	50	50
193	36	46	6	70	31
202	37	62	12	210	120
176	37	54	4	60	25
157	32	52	11	230	80
156	33	54	15	225	73
138	33	68	2	110	43

编程法：

编写如下程序（其在光盘中的存储路径为“proc\chap11\fit”）：

```
proc cancorr data=chap11.Fit all
        vprefix=Physiological vname='Physiological Measurements'
        wprefix=Exercises wname='Exercises';
 /*调用 cancorr 过程，并指定输出相关系数、典型相关检验 P 值等统计量,
 分别指定 var 和 with 变量的前缀和变量组名*/
    var Weight Waist Pulse;
    with Chins Situps Jumps;
 /*分析变量组 Weight,Waist,Pulse 和变量组 Chins,Situps,Jumps 之间的典型相关关系*/
run;
```

选择 Run|Submit 命令提交程序，以下分析输出的主要结果。

1．典型相关系数及其检验结果

表 11-7 为典型相关系数的信息。此表第一行为第一对典型相关变量（Physiological1，Exercises1）之间典型相关系数的信息：典型相关系数（Canonical Correlation）为 0.795608，校正典型相关系数（Adjusted Canonical Correlation）为 0.754056，近似标准误（Approximate Standard Error）为 0.084197，典型相关系数的平方（Squared Canonical Correlation）为 0.632992，以及 $r_1^2/(1-r_1^2)$ 对应的特征值（Eigenvalue）、两相邻特征值的差（Difference）、特征值的贡献率（Proportion）和累计贡献率（Cumulative）的值。第一个特征值的贡献率达到了 97.34%，表格第二、第三行分别为第二、第三对典型相关变量之间的典型相关系数信息。

表 11-7　典型相关系数信息表

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)			
					Eigenvalue	Difference	Proportion	Cumulative
1	0.795608	0.754056	0.084197	0.632992	1.7247	1.6828	0.9734	0.9734
2	0.200556	−0.076399	0.220188	0.040223	0.0419	0.0366	0.0237	0.9970
3	0.072570		0.228208	0.005266	0.0053		0.0030	1.0000

表 11-8 为对典型相关系数 r 的 F 检验结果，对第一典型相关系数 r_1 检验的原假设为“r_1 及所有小于 r_1 的典型相关系数的值为零”，对第二典型相关系数检验的原假设为“r_2 及所有小于 r_2 的典型相关系数的值为零”，依此类推。在本实验中，r_1 对应的 F 检验 P 值为 0.0635，小于显著性水平 0.1；而观察到其他典型变量都没有显著的统计意义，不予考虑。由表 11-9 对 r_1 的多元统计 Wilk’s Lambda 检验结果，在显著性水平为 0.1 的水平下拒绝“r_1 及所有小于 r_1 的典型相关系数的值为零”的原假设。

表 11-8　典型相关系数 F 检验结果

	Test of H0: The canonical correlations in the current row and all that follow are zero				
	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.35039053	2.05	9	34.223	0.0635

续表

	Test of H0: The canonical correlations in the current row and all that follow are zero				
	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
2	0.95472266	0.18	4	30	0.9491
3	0.99473355	0.08	1	16	0.7748

表 11-9　多元统计检验结果

Multivariate Statistics and F Approximations					
S=3　M= -0.5　N=6					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilk's Lambda	0.35039053	2.05	9	34.223	0.0635
Pillai's Trace	0.67848151	1.56	9	48	0.1551
Hotelling-Lawley Trace	1.77194146	2.64	9	19.053	0.0357
Roy's Greatest Root	1.72473874	9.20	3	16	0.0009
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					

2．原始的典型系数

表 11-10 为生理指标（Physiological）变量的原始典型系数，表 11-11 为训练指标（Exercises）变量的原始典型系数。据此可写第一对典型变量（Physiological1,Exercises1）和原始变量的线性方程：

Physiological1=−0.031404688* Weight +0.4932416756* Waist−0.008199315* Pulse

Exercises1=−0.066113986* Chins−0.016846231* Situps+0.0139715689* Jumps

表 11-10　生理指标的原始典型系数

Raw Canonical Coefficients for the Physiological Measurements			
	Physiological1	Physiological2	Physiological3
Weight	−0.031404688	−0.076319506	−0.007735047
Waist	0.4932416756	0.3687229894	0.1580336471
Pulse	−0.008199315	−0.032051994	0.1457322421

表 11-11　训练指标的原始典型系数

Raw Canonical Coefficients for the Exercises			
	Exercises1	Exercises2	Exercises3
Chins	−0.066113986	−0.071041211	−0.245275347
Situps	−0.016846231	0.0019737454	0.0197676373
Jumps	0.0139715689	0.0207141063	−0.008167472

3．标准化的典型系数

由于各个分析变量的单位并不一致，为了降低量纲不同而导致的结果偏误，表 11-12 和表 11-13

列出了标准化的典型系数，据此可写出典型变量（Physiological1，Exercises1）和标准化变量的线性方程：

Physiological1=−0.7754* Weight +1.5793* Waist−0.0591* Pulse

Exercises1=−0.3495* Chins−1.0540* Situps+0.7164* Jumps

表 11-12　体能指标标准化典型系数

Standardized Canonical Coefficients for the Physiological Measurements			
	Physiological1	Physiological2	Physiological3
Weight	−0.7754	−1.8844	−0.1910
Waist	1.5793	1.1806	0.5060
Pulse	−0.0591	−0.2311	1.0508

表 11-13　训练指标标准化典型系数

Standardized Canonical Coefficients for the Exercises			
	Exercises1	Exercises2	Exercises3
Chins	−0.3495	−0.3755	−1.2966
Situps	−1.0540	0.1235	1.2368
Jumps	0.7164	1.0622	−0.4188

4．典型结构

表 11-14 至表 11-17 为 4 个典型结构矩阵，表 11-14 为生理指标原始变量组（Weight、Waist 和 Pulse）和它对应的三个典型变量之间的相关系数矩阵；表 11-15 为训练指标原始变量组（Chins、Situps 和 Jumps）和它对应的三个典型变量之间的相关系数矩阵；表 11-16 为生理指标原始变量组和训练指标三个典型变量之间的相关系数矩阵；表 11-17 为训练指标原始变量组和生理指标三个典型变量之间的相关系数矩阵。

表 11-14　生理指标变量组与其构成变量的相关系数矩阵

Correlations Between the Physiological Measurements and Their Canonical Variables			
	Physiological1	Physiological2	Physiological3
Weight	0.6206	−0.7724	−0.1350
Waist	0.9254	−0.3777	−0.0310
Pulse	−0.3328	0.0415	0.9421

表 11-15　训练指标变量组与其构成变量的相关系数矩阵

Correlations Between the Exercises and Their Canonical Variables			
	Exercises1	Exercises2	Exercises3
Chins	−0.7276	0.2370	−0.6438
Situps	−0.8177	0.5730	0.0544
Jumps	−0.1622	0.9586	−0.2339

表 11-16　生理指标变量组和训练指标典型变量的相关系数矩阵

Correlations Between the Physiological Measurements and the Canonical Variables of the Exercises			
	Exercises1	Exercises2	Exercises3
Weight	0.4938	−0.1549	−0.0098
Waist	0.7363	−0.0757	−0.0022
Pulse	−0.2648	0.0083	0.0684

表 11-17　训练指标变量组和生理指标典型变量的相关系数矩阵

Correlations Between the Exercises and the Canonical Variables of the Physiological Measurements			
	Physiological1	Physiological2	Physiological3
Chins	−0.5789	0.0475	−0.0467
Situps	−0.6506	0.1149	0.0040
Jumps	−0.1290	0.1923	−0.0170

表 11-18 和表 11-19 分别为原始生理指标变量和训练指标变量被它们自己的典型相关变量和对方的典型相关变量解释的方差比。例如，生理指标变量组被典型变量 Physiological1 解释了 37.12%的方差，被典型变量 Physiological2 解释了 54.36%的方差；而被典型变量 Exercises1 解释了 23.49%的方差，被典型变量 Exercises2 解释了 2.19%的方差。

表 11-18　原始生理指标变量被典型变量所解释的方差比

Raw Variance of the Physiological Measurements Explained by					
Canonical Variable Number	Their Own Canonical Variables		Canonical R-Square	The Opposite Canonical Variables	
	Proportion	Cumulative Proportion		Proportion	Cumulative Proportion
1	0.3712	0.3712	0.6330	0.2349	0.2349
2	0.5436	0.9148	0.0402	0.0219	0.2568
3	0.0852	1.0000	0.0053	0.0004	0.2573

表 11-19　原始训练指标变量被典型变量所解释的方差比

Raw Variance of the Exercises Explained by					
Canonical Variable Number	Their Own Canonical Variables		Canonical R-Square	The Opposite Canonical Variables	
	Proportion	Cumulative Proportion		Proportion	Cumulative Proportion
1	0.4111	0.4111	0.6330	0.2602	0.2602
2	0.5635	0.9746	0.0402	0.0227	0.2829
3	0.0254	1.0000	0.0053	0.0001	0.2830

类似的，表 11-20 和表 11-21 分别为标准化的生理指标变量和训练指标变量被它们自己的典

型相关变量和对方的典型相关变量解释的方差比。例如，标准化的生理指标变量组被典型变量 Physiological1 解释了 45.08%的方差，被典型变量 Physiological2 解释了 24.7%的方差；而被典型变量 Exercises1 解释了 28.54%的方差，被典型变量 Exercises2 解释了 0.99%的方差。

表 11-20　标准化的生理指标变量被典型变量所解释的方差比

Standardized Variance of the Physiological Measurements Explained by					
Canonical Variable Number	Their Own Canonical Variables		Canonical R-Square	The Opposite Canonical Variables	
	Proportion	Cumulative Proportion		Proportion	Cumulative Proportion
1	0.4508	0.4508	0.6330	0.2854	0.2854
2	0.2470	0.6978	0.0402	0.0099	0.2953
3	0.3022	1.0000	0.0053	0.0016	0.2969

表 11-21　标准化的训练指标变量被典型变量所解释的方差比

Standardized Variance of the Exercises Explained by					
Canonical Variable Number	Their Own Canonical Variables		Canonical R-Square	The Opposite Canonical Variables	
	Proportion	Cumulative Proportion		Proportion	Cumulative Proportion
1	0.4081	0.4081	0.6330	0.2584	0.2584
2	0.4345	0.8426	0.0402	0.0175	0.2758
3	0.1574	1.0000	0.0053	0.0008	0.2767

以下结合本例的背景资料进行分析，由典型相关系数检验结果显示，仅仅第一对典型相关系数达到了统计显著，因此只分析第一对典型变量的相关情况：生理指标的第一个典型变量 Physiological1 在标准化的 Weight（体重）上的典型系数为-0.7754，在标准化 Waist（胸围）上的典型系数为 1.5793，而在标准化变量 Pulse（心跳次数）上的典型系数为负值且较小，即说明典型变量 Physiological1 主要代表了健身人员的腰围和体重的信息；训练指标的第一个典型变量 Exercises1 在标准化变量 Situps（仰卧起坐）的典型系数为负且最大（-1.0540），在标准化变量 Jumps（跳高）上的典型系数为 0.7164，说明健身人员的体重和腰围对他们仰卧起坐和跳高的成绩影响较大。

菜单法：

步骤一：选择 Solutions|Analysis|Analyst 命令，进入 Analyst 模块主界面。

步骤二：选择 File|Open By SAS Name | Chap11|fit |OK 命令，打开数据集 chap11.fit。

步骤三：选择 Statistics|Multivariate|Canoncial Correlation 命令，显示典型相关分析主对话框，如图 11-1 所示。按住鼠标拖动选择变量 Weight、Waist 和 Pulse，单击 Set1（第一组）按钮，将这些变量定义为第一组变量集合。类似的，将变量 Chins、Jumps 和 Situps 定义为第二组变量集合（Set2）。

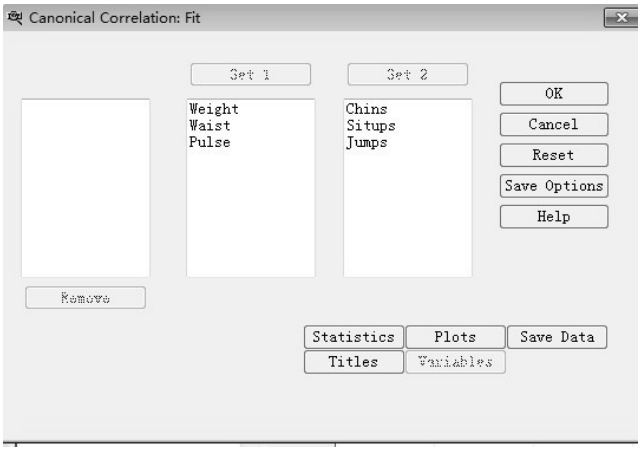


图 11-1　典型相关分析主对话框

步骤四：单击 Statistics（统计量）按钮，弹出如图 11-2 所示对话框，在选项# of canonical variables（典型变量个数）选项后填入 2。可分别在 Set 1 canonical variables（第一组典型变量）和 Set 2 canonical variables（第二组典型变量）选项框内设置第一组和第二组典型变量的 Label（标签）和 Prefix（前缀）。单击 OK 按钮返回如图 11-1 所示对话框。

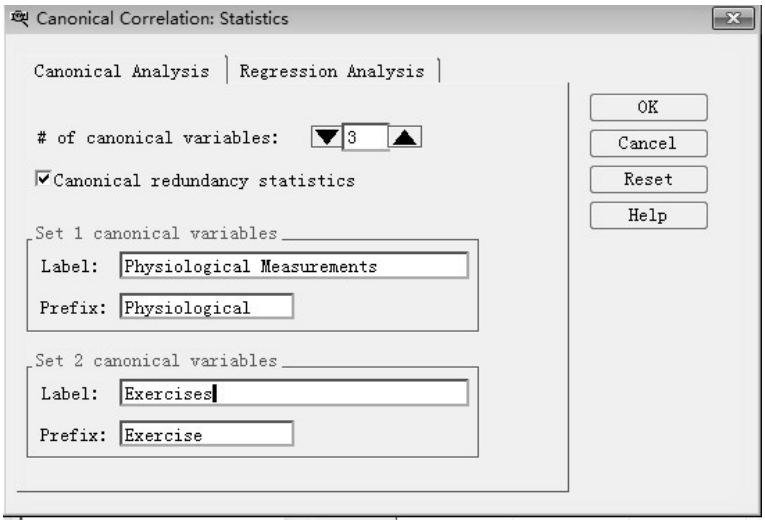


图 11-2　设置典型相关分析

步骤五：单击 Plots（绘图）按钮，弹出如图 11-3 所示对话框，选择 Create canonical variable plots（绘制典型变量图），Canonical variables（典型变量）选项采用系统默认为“1”to“2”，即绘制典型变量 1 和典型变量 2 的图形。单击 OK 按钮保存设置并返回如图 11-1 所示对话框。单击 OK 按钮，则系统将输出和编程方法类似的结果。

在图 11-1 所示的对话框单击 Save Data 按钮，可在弹出的对话框中进行保存得分数据集和统计量数据集；单击 Titles 按钮，在弹出的对话框中进行标题设置；单击 Variables 按钮，在弹出的对话框中可设置分层变量、加权变量和频数变量。

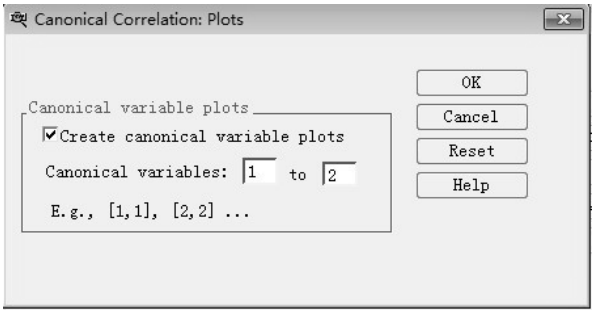


图 11-3　设置典型变量图

练习题

习题　为了探索雇员的工作满意度和工作特征之间的关系，某研究人员调查了 14 名雇员，并记录下他们衡量工作满意度的工作前景和发展满意度（Career）、对领导者和管理者的管理方式满意度（Supervisor）、对薪水的满意度（Finance）三个指标和工作特征的三个指标，即工作任务类型（Variety）、灵活度（Feedback）和自主权（Autonomy）三个指标的评分，表 11-12 为调查数据（相应的 SAS 数据集在光盘中的存储路径为“data\chap11\jobs”）。请分析雇员工作特征和工作满意度之间的相关关系。

表 11-22　雇员工作特征和工作满意度调查数据

Career	Supervisor	Finance	Variety	Feedback	Autonomy
72	26	9	10	11	70
63	76	7	85	22	93
96	31	7	83	63	73
96	98	6	82	75	97
84	94	6	36	77	97
66	10	5	28	24	75
31	40	9	64	23	75
45	14	2	19	15	50
42	18	6	33	13	70
79	74	4	23	14	90
39	12	2	37	13	70
54	35	3	23	74	53
60	75	5	45	58	83
63	45	5	22	67	53

（本习题的解答程序在光盘中的存储路径为“proc\chap11\jobs”。）

第12章　聚 类 分 析

聚类分析，顾名思义是研究“物以类聚”的一种多元统计分析方法，把分类对象按一定规则分成组或类，这些组或类不是事先给定的而是根据数据特征而定的。每一类中包含的对象差异小，而类间对象的差异较大。例如，根据食物中各个营养成分的含量将其分成不同类别，根据客户的消费特性将其分成不同的群体等都属于聚类分析问题。

本章首先介绍聚类分析的基本原理，然后介绍主要用于样品聚类的 CLUSTER 过程和用于变量聚类的 VARCLUS 过程，最后结合实例介绍应用 SAS 系统实现聚类分析。

12.1　基本原理

本节主要介绍聚类分析的基本原理，首先介绍聚类分析的最小单位——样品（变量）间距离系数的定义；然后介绍类的性质，包括类的定义、三个特征指标和类间距离；最后介绍常用的聚类方法及决定聚类次数的统计量。

12.1.1　样品（变量）间距离定义

聚类分析的目的是归类“相似”的对象，对象主要有样品和变量。相应的有两类衡量“相似”的聚类统计量：一类是主要用于样品聚类的统计指标——类与类之间距离，即把每个样品当作高维空间中的一个点，类与类之间按照某种准则规定它们的距离，将距离近的点聚合成一类，距离远的点聚合成另一类；另一类是主要用于变量聚类的相似系数，根据相似系数将差异小的变量归为一类，把差异大的变量归为另一类。以下介绍常见的距离和相似系数。设有 n 组样品，每个样品对应 p 个变量，如表 12-1 所示。

表 12-1　p 个变量的 n 组样品数据

变量 / 样品	X_1　X_2 … X_p
1	x_{11}　x_{12}　…　x_{1p}
2	x_{21}　x_{22}　…　x_{2p}
⋮	⋮　⋮　⋮　⋮
n	x_{n1}　x_{n2}　…　x_{np}

用于样品聚类的第 i 个与第 j 个样品之间的距离用 d_{ij} 表示，d_{ij} 应满足以下条件：

- 当第 i 个样品与第 j 个样品相等时 $d_{ij}=0$；
- 对一切 i、j 有 $d_{ij}\geqslant 0$；

- 对一切 i、j 有 $d_{ij}=d_{ji}$；
- 对一切 i、j，k 有 $d_{ij}\leqslant d_{ik}+d_{kj}$。

最常用的距离有欧几里得距离、闵可夫斯基距离和马氏距离，它们的具体定义如下：

- 欧几里得（Euclid）距离：$d_{ij}=\left(\sum_{k=1}^{p}\left(x_{ik}-x_{jk}\right)^2\right)^{\frac{1}{2}}$
- 闵可夫斯基（Minkowski）距离：$d_{ij}=\left(\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|^g\right)^{1/g}$，$g$ 一般为 1 或 2，如果 g=1 时也称为绝对值距离，g=2 时即为欧几里得距离。
- 马哈那诺比斯（Mathalanobis）距离：$d_{ij}=(\boldsymbol{x}_i-\boldsymbol{x}_j)'\boldsymbol{S}^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_j)$，其中 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 分别为第 i、j 个样品的 p 个变量组成的向量，$\boldsymbol{S}^{-1}$ 为 n 个样品的 $p\times p$ 的协方差矩阵的逆矩阵。

对变量的聚类分析通常基与变量之间的相似系数定义变量间的距离，相似系数常用夹角余弦和相关系数来衡量。

- 夹角余弦：记变量 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 的夹角余弦为 c_{ij}，其中 $i,j=1,2,\cdots,p$，则有：

$$c_{ij}=\frac{\sum_{k=1}^{n}x_{ik}x_{jk}}{\left(\sum_{k=1}^{n}x_{ik}^{\,2}\sum_{k=1}^{n}x_{jk}^{\,2}\right)^{1/2}}$$

- 相关系数：变量 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 的相关系数为 $r_{ij}=\dfrac{\sum_{k=1}^{n}(x_{ik}-\overline{x}_i)(x_{jk}-\overline{x}_j)}{\left[\sum_{k=1}^{n}(x_{ik}-\overline{x}_i)^2\sum_{k=1}^{n}(x_{jk}-\overline{x}_j)^2\right]^{\frac{1}{2}}}$，$\overline{x}_i$ 表示第 i 个指标的平均值。

于是定义 $d_{ij}=1-c_{ij}^2$ 或 $d_{ij}=1-r_{ij}^2$ 为变量之间的距离。

12.1.2　类的性质

本章分析的是聚类，那到底什么是“类”呢？以下给出类的三种定义。用 G 表示类，假设 G 中有 k 个元素，用 i、j 表示 G 中第 i 个、第 j 个因素，d_{ij} 为 i 和 j 的距离，T 为一给定的阈值。

定义 1　若对任意的 $i,j\in G$ 有 $d_{ij}\leqslant T$，则称 G 为一个类。

定义 2　若对每个 $i\in G$ 有 $\dfrac{1}{k-1}\sum_{j\in G}d_{ij}\leqslant T$，则称 G 为一个类。

定义 3　若对任意一个 $i\in G$，一定存在 $j\in G$，使得 $d_{ij}\leqslant T$，则称 G 为一个类。

常用以下三种方式描绘类的特征，已知类 G 的元素用 $x_1,\cdots,x_m$ 表示，m 为 G 内的样品数。

均值 $\overline{x}_G$（或称为 G 的重心）：$\overline{x}_G=\dfrac{1}{m}\sum_{i=1}^{m}x_i$

样品协方差矩阵：$\boldsymbol{S}_G = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x}_G)(x_i - \overline{x}_G)'$

G 的直径：定义不唯一，如 $D_G = \max_{i,j\in G} d_{ij}$，$D_G = \sum_{i\in G}(x_i - \overline{x}_G)'(x_i - \overline{x}_G)$。

在聚类分析中，不仅要考虑各个类的特征，还要计算类与类之间的距离。由于类的形状是多种多样的，所以类与类之间的距离也有多种计算方法。令 G_p 和 G_q 中分别有 p 和 q 个样品，它们的重心分别记为 $\overline{x}_p$ 和 $\overline{x}_q$。以下列出一些常用的类与类之间距离 $D(p,q)$ 的定义：

- 最短距离：$D(p,q) = \min\left\{d_{jk} \middle| j \in G_p, k \in G_q\right\}$
- 最长距离：$D(p,q) = \max\left\{d_{jk} \middle| j \in G_p, k \in G_q\right\}$
- 重心法距离：$D(p,q) = (\overline{x}_p - \overline{x}_q)'(\overline{x}_p - \overline{x}_q)$
- 类平均距离：$D(p,q) = \frac{1}{pq}\sum_{i\in G_p}\sum_{j\in G_q} d_{ij}$
- Ward 离差平方和距离：用 D_p、D_q 分别表示 G_p 和 G_q 的直径，用 D_{p+q} 表示类 $D_p \cup D_q$ 的直径，即 $D_p = \sum_{i\in G_p}(x_i - \overline{x}_p')(x_i - \overline{x}_p)$，$D_q = \sum_{i\in G_q}(x_j - \overline{x}_q)'(x_j - \overline{x}_q)$，$D_{p+q} = \sum_{i\in G_p\cup G_q}(x_i - \overline{x})'(x_i - \overline{x})$。其中 $\overline{x} = \frac{1}{p+q}\sum_{i\in G_p\cup G_q} x_i$。用离差平方和法定义 G_p 和 G_q 之间的距离即 $D_w(p,q) = D_{p+q} - D_p - D_q$。
- 密度估计法：密度估计法是一类使用非参数概率密度的聚类方法，包括两个步骤：首先使用一种基于密度估计的新的非相似测度 d^* 来计算样品 x_i 和 x_j 的近邻关系；然后根据基于 d^* 方法计算的距离，采用最小距离法进行聚类。有 5 种不同的密度估计法：k 最近邻估计法、均匀核估计、Wong 混合法、两阶段密度估计法和最大似然估计法。

12.1.3 聚类方法

以下介绍几种常见的聚类方法。

（1）系统聚类法：目前最常用的方法，基本思想是首先将 n 个样品看成 n 类，然后指定样品间的距离和类与类之间的距离。将距离最近的两类合并为一个新类，再计算新类和其他类之间的距离，从中找出最近的两类合并，注意如果在某一步的距离矩阵中最小元素不止一个时可以将其同时合并继续下去，直到最后所有的样品被归为一类。

（2）动态聚类法：开始将 n 个样品粗略地分成若干类，然后用某种最优准则进行调整，一次又一次地调整，直至不能调整为止。

（3）分解法：与系统聚类相反，开始时所有的样品都在一类，然后用某种最优准则将它分成两类，再用同样准则将这两类各自试图分裂为两类，从中选出一个使目标函数较好者，这样由两类变成了三类。如此下去，一直分裂到每类只有一个样品为止（或用其他停止规则）。

（4）加入法：将样品依次输入，每次输入后将它放到当前聚类图的应有位置上，全部输入后，即得聚类图。

聚类过程经常通过计算、观察类的各种统计量来决定聚类个数，类的统计量主要有：

（1）类的均方根标准差：计算公式为：

$$\text{RMSSTD}=\sqrt{D_G/(v(p-1))}$$

其中，D_G 为类 G_p 的直径，即类内的离差平方和。可以形象理解成如果一个类的离差平方和等于 0，那么类内的所有点都集中在一个点上，则类的直径为 0；如果一个类的离差平方和逐渐变大，那么类内的所有点就越来越分散，包含所有点的一个圆或球就会越来越大，相应地这个圆或球的直径就越来越大。v 为样品包括的变量个数，p 为类中包含的样品个数。

（2）R^2 统计量：计算公式为：

$$R^2=1-\sum D_i/\text{TSS}$$

其中，$\sum D_i$ 为在第 i 次分类时对 G 个类的直径求和，TSS 为所有样品的总离差平方和。一般来说，R^2 统计量用于评价每次合并成 i 个类时的聚类效果。当 $\sum D_i/\text{TSS}$ 值越小（也即 R^2 统计量越大，越接近 1），表示类内离差平方和 $\sum D_i$ 在总离差平方和 TSS 中所占的比例越小，说明了这 G 个类越分开，故聚类效果越好。R^2 的值总是在 0 和 1 之间，当 n 个样品各自为一类时，$R^2=1$，说明类被完全分开；当 n 个样品最后合并成一类时，$R^2=0$，说明类被完全混合在一起了，分不清楚了。而且 R^2 的值总是随着分类个数的减少而变小。如何根据 R^2 的值来确定 n 个样品被分成多少类呢？首先，最合适分类的 R^2 的值不能太小，最好能达到 0.7 以上；其次，观察 R^2 值的变化，如果某次合并使 R^2 值减小很多，则最佳合并次数为当次合并数减一。

（3）半偏 R^2 统计量：合并类 G_p 和类 G_q 为类 G_m 时，可以用半偏 R^2 统计量评价此次合并的效果，其计算公式为：

$$\text{半偏}R^2=D_w(p,q)/\text{TSS}$$

其中，$D_w(p,q)$ 表示合并类 G_p 和类 G_q 为新类 G_m 后，类内离差平方和的增量。显然，半偏 R^2 值=上次合并后 R^2 值－这次合并后 R^2 值，则半偏 R^2 值较大的上次合并为最佳合并。

（4）伪 F 统计量：计算公式为：

$$\text{伪}F(v(G-1),v(n-G))=\frac{(\text{TSS}-\sum D_i)/(G-1)}{\sum D_i/(n-G)}$$

其中，G 为聚类的个数；n 为样品总数；v 为样品包含的变量数；自由度为 $v(G-1)$ 和 $v(n-G)$。在给定显著性水平上该统计量用于评价分为 G 个聚类的效果。伪 F 值越大越表示这些观察可显著地分为 G 个类。

（5）伪 t^2 统计量：计算公式为：

$$\text{伪}t^2(v,v(p+q-2))=\frac{D_m-D_p-D_q}{(D_p+D_q)/(p+q-2)}$$

该统计量用以评价合并类 G_p 和类 G_q 的效果。该值很大说明合并类 G_p 和类 G_q 为类 G_m 后，使得离差平方和的增加量（$D_m-D_p-D_q$）相对于原来 G_p 和 G_q 两类的类内离差平方和很大。这表明上一次合并的两个类 G_p 和 G_q 是分开得很大的，即上一次聚类的效果较好。伪 t^2 和伪 F 统计量都可以作为确定类个数的有效指标。

12.2 样品聚类

12.2.1 SAS 过程——CLUSTER 过程

CLUSTER 过程主要应用最短距离法、最长距离法、平均距离法等参数方法和密度估计法、两阶段密度估计法等非参数方法对样品进行聚类。

CLUSTER 过程一般使用格式如下：

```
PROC CLUSTER    <选项列表>;
VAR          变量列表;
ID           变量;
FREQ         变量;
COPY         变量表;
RMSSTD       变量;
BY           变量表;
RUN;
```

PROC CLUSTER 语句后主要的控制选项为：输入、输出数据集选项（如表 12-2 所示），聚类方法细节选项（如表 12-3 所示），打印输出选项（如表 12-4 所示）。

表 12-2 输入、输出数据集选项

选　项	意　义
DATA=SAS 数据集	定义包含进行聚类的观测的输入数据集
OUTTREE=SAS 数据集	新建一个供 TREE 过程调用输出聚类结果的树状图的数据集

表 12-3 聚类方法细节选项

选　项	意　义
METHOD=算法	指定聚类方法： WARD（离差平方和法）、AVERAGE（类平均法）、CENTROID（重心法）、COMPLETE（最长距离法）、SINGLE（最短距离法）、MEDIAN（中间距离法）、DENSITY（密度法）、FLEXIBLE（可变类平均法）、TWOSTAGE（两阶段密度法）、EML（最大似然法）、MCQUITTY（相似分析法）
STANDARD	对变量实施标准化
NONORM	阻止距离被正态化成均值为 1 或均方根为 1
NOSQUARE	阻止 CANDISC 过程在 METHOD= AVERAGE、CENTROID、MEDIAN、WARD 方法中将距离数据平方
MODE=N	当合并两个类时，规定对被指定的众数类中的每类至少有 N 个成员。此选项只能与 METHOD= DENSITY 或 TWOSTAGE 联用
TRIM=P	要求从分析中删去那些概率密度估计较小的点。P 的有效值为 0≤P<100，被当作百分比。在使用 METHOD= WARD 或 COMPLETE 时，聚类可能会被异常值严重歪曲，因此最好使用这个选项。此选项也可用于 METHOD= SINGLE 中
DIM=N	当规定 METHOD= DENSITY 或 TWOSTAGE 时指定使用的维数。N 值必须设置为大于或等于 1。如果数据是坐标数据，默认值为变量的个数；如果是距离数据，默认值为 1

续表

选　项	意　义
HYBRID	要求用 WONG 混合聚类方法，其中密度用 K 均值法的初始聚类分析中的均值计算得到。这个选项只能在规定 METHOD= DENSITY 或 TWOSTAGE 时使用
K=N	指明 K 最近邻估计法中近邻的个数。近邻个数 N 必须大于或等于 2 且小于观察数
R=N	指明均匀核密度估计法的支撑球半径。N 值必须设置为大于 0
NOTIE	阻止 CLUSTER 过程在聚类历史过程中检查每次产生的类间最小距离连接（TIES）的情况。规定这个选项以便减少过程执行的时间和空间

表 12-4　打印输出选项

选　项	意　义
RSQUARE	输出 R^2 和半偏 R^2
RMSSTD	输出每一类的均方根标准差
CCC	输出在均匀的原假设下判断聚类分成几类合适的一种立方聚类准则统计量 CCC 和近似期望值 R^2。同时打印输出选项 RSQUARE 有关的 R^2 和半偏 R^2。此选项不适合于 METHOD=SINGLE，因为该方法容易删掉分布的结尾部分
PSEUDO	输出伪 F 统计量（标志为 PSF）和伪 t^2 统计量（标志为 PST2）。当分类数目不同时，它们有不同的取值
SIMPLE	打印简单统计量
STD	标准化变量

CLUSTER 过程中使用的其他语句含义如下：

COPY 语句——指明输入数据集中的一些变量复制到 OUTTREE=的输出数据集中。

RMSSTD 语句——当输入数据集中的坐标数据代表类的均值时，定义表示均方根标准差变量，通常与 FREQ 语句中的变量配合使用。

12.2.2　SAS 过程——TREE 过程

本过程利用样品聚类 CLUSTER 过程和变量 VARCLUS 过程生成的数据集来绘制树状结构图。这个树状结构图可以按垂直或水平方向输出。TREE 过程可以把输入数据集中的任何数值变量都能够用来规定这些类的高度，还可根据用户的要求生成一个输出数据集，其中包含一个变量，其值用以标志在这个树里指定水平上不相交的类。

TREE 过程的一般使用格式如下：

```
PROC TREE      <选项列表>;
NAME        变量;
PARENT      变量;
HEIGHT      变量;
ID          变量;
COPY        变量列表;
FREQ        变量;
BY          变量列表;
RUN;
```

PROC TREE 语句后可使用的控制选项按性质分为以下 4 类：输入、输出数据集选项（如表 12-5 所示），树状结构控制选项（如表 12-6 所示），树高度和树叶控制选项（如表 12-7 所示）和其他选项（如表 12-8 所示）。

表 12-5　输出、输出数据集选项

选　　项	意　　义
DATA=SAS 数据集	输入由 CLUSTER 过程和 VARCLUS 过程生成的数据集
OUT=SAS 数据集	新建一个包括绘制树形结构图的有关数据的数据集

表 12-6　控制树状结构选项

选　　项	意　　义
LEVEL =N	规定确定不相交类的树状图水平（层次）
NCL =N	指定希望聚类个数
DOCK =N	当某个类中的对象（观察或变量）的个数小于或等于 N 时，在输出数据集里把该类中这些对象的变量 CLUSTER 和 CLUSNAME 的值设置为缺失值。系统默认 N 为 0
ROOT =“名称值”	若不想输出整个树状图，规定想输出的子树根的 NAME 变量的值
HOR	要求树状图的取向为水平方向，且树根在左边。若未指明此选项，则其为垂直方向，树根在上部

表 12-7　控制树高度和树叶控制选项

选　　项	意　　义
HEIGHT=	规定在树状图中用以确定高度轴的常规变量，可设置为 HEIGHT=H/L/M/N，分别代表_HEIGHT_变量、根到自己节点的路径长度、_MODE_变量、_NCL_变量
MAXH=N	指定在高度轴上打印的最大值
MINH=N	指定在高度轴上打印的最小值
NTICH=N	指定在高度轴上刻度之间的间隔个数
PAGES=N	规定将此树状图展开的页数
POS=N	指定在高度轴上打印位置的个数
SPACES=N	规定在打印输出中对象之间的空格数
TICKPOS=N	指定在高度轴上每个刻度间隔打印位置的个数
FILLCHAR=“字母”	规定没有连成一类的树叶之间的打印字符，默认值为空格
JOINCHAR=“字母”	规定已连成一类的树叶之间的打印字符，默认值为 X
LEAFCHAR=“字母”	规定表示没有子辈的类的打印字符，默认值为“.”
TREECHAR=“字母”	规定表示有子辈的类的打印字符，默认值为“X”

表 12-8　其他选项

选　　项	设　　置
SORT	按照聚类的形成顺序，用 HEIGHT 变量对每个节点的子辈排序
DES	把选项 SORT 的排列顺序反过来
LIST	列出这个树中所有节点，并且打印高度、父辈及每个节点的子辈

续表

选　项	设　置
NOPRINT	创建 OUT=的输出数据集而不绘制树状图
GRAPHICS	要求在 GRAPH 窗口中输出高分辨率的树状图，类的合并用连接线归纳表示

TREE 过程中使用的其他语句含义如下：

NAME 语句——规定一个用以标志每个观察代表的节点的字符或数值变量。NAME 变量同 PARENT 变量联合确定树的结构。语句默认时寻找_NAME_变量。

PARENT 语句——规定一个标志每个观察的父辈节点的字符或数值变量。语句默认时寻找_PARENT_变量。

HEIGHT 语句——规定一个用于定义这个树中每个节点（类）的高度数值变量。高度变量由选项 HEIGHT=规定。

COPY 语句——把语句中列出的一个或几个变量复制到 OUT=的输出数据集中。

ID 语句——ID 变量可以是字符或数值变量，用以在打印输出树状图中识别对象。

12.2.3　SAS 实例——根据飞行距离对 10 所美国城市分类

例 12-1　已知美国 10 所城市之间的飞行距离数据，如表 12-9 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap12\ mileages”）。请根据城市间的距离对它们进行分类。

表 12-9　10 所城市飞行距离数据

City	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Atlanta	0									
Chicago	587	0								
Denver	1212	920	0							
Houston	701	940	879	0						
Los Angeles	1936	1745	831	1374	0					
Miami	604	1188	1726	968	2339	0				
New York	748	713	1631	1420	2451	1092	0			
San Francisco	2139	1858	949	1645	347	2594	2571	0		
Seattle	2182	1737	1021	1891	959	2734	2408	678	0	
Washington D.C.	543	597	1494	1220	2300	923	205	2442	2329	0

注意：变量 C1 至 C10 分别代表城市：Atlanta，Chicago，Denver，Houston，Los Angeles，Miami，New York，San Francisco，Seattle，Washington D.C。

编程法：

编写如下程序（其在光盘中的存储路径为“proc\chap12\ mileages”）：

```
ods graphics on;
proc cluster data=chap12.mileages        /*调用 cluster 过程*/
              outtree=tree               /*输出数据集 tree 以绘制树状图*/
              method=average             /*应用类平均法进行聚类分析*/
```

```
                    pseudo;                    /*要求输出伪 t2 和伪 F 统计量*/
id City;                                       /*定义标识变量为 city*/
run;
proc tree data=tree HOR ; /*调用 tree 过程绘制树状图，且指定树根在左侧*/
id City;
run;
ods graphics off;
```

选择 Run|Submit 命令提交程序，以下分析平均距离法进行样品聚类分析的主要输出结果。

表 12-10 为 10 个样品依次聚成 10～1 类的过程及结果。以下详细解释结果：NCL 列为聚类数；Clusters Joined 列为每次聚成一个新类的两个样品（即为城市名，如 Los Angeles）或旧类（如 CL8）；FREQ 列为新类中含有的样品数，如 NCL 值为 5 时，将旧类 CL8（包括 Los Angeles 和 San Francisco）和新样品“Seattle”聚成新类，则其对应的 FREQ 值为 3；PSF 和 PST2 分别为伪 F 统计量和伪 t^2 统计量。当 PST2 出现峰值的前一类所对应的分类数较合适。图 12-1 所示的横轴为聚类个数，纵轴为对应的 PSF 和 PST2 的值，观察可得当聚合为 5 类时，伪 t^2 统计量取最大值，则聚为 4 类比较合适。

Norm RMS Dist 列为两样品或两类间的平均距离。Tie 列为结的个数，在本例中未出现打结的情况。

打结：在系统聚类的每一层，CLUSTER 必须按最小距离把两类合并，对于离散距离而言，偶尔会出现几个相等的最小距离，此时便出现了打结的情况。类是采用内观测最小的序号来识别的，这两类有一个较大序号和一个较小序号，若出现结，则取其中较大序号中的最小者合并，在输出的 Tie 列中，以 T 指出最小距离的一个结，空白表明没有结。

表 12-10　聚类过程

Cluster History							
NCL	Clusters Joined		FREQ	PSF	PST2	Norm RMS Dist	Tie
9	New York	Washington D.C.	2	66.7	.	0.1297	
8	Los Angeles	San Francisco	2	39.2	.	0.2196	
7	Atlanta	Chicago	2	21.7	.	0.3715	
6	CL7	CL9	4	14.5	3.4	0.4149	
5	CL8	Seattle	3	12.4	7.3	0.5255	
4	Denver	Houston	2	13.9	.	0.5562	
3	CL6	Miami	5	15.5	3.8	0.6185	
2	CL3	CL4	7	16.0	5.3	0.8005	
1	CL2	CL5	10	.	16.0	1.2967	

图 12-2 为树状图，纵轴为城市名，横轴代表类间的平均距离。根据笔者在图上添加的直线，将城市聚成 4 类，Atlanta、Chicago、New York、Washington D.C 为一类，Miami 为一类，Denver 和 Houston 为一类，Los Angeles、San Francisco 和 Seattle 为一类。即同类的城市之间飞行距离较近，而不同类的城市之间飞行距离较远。

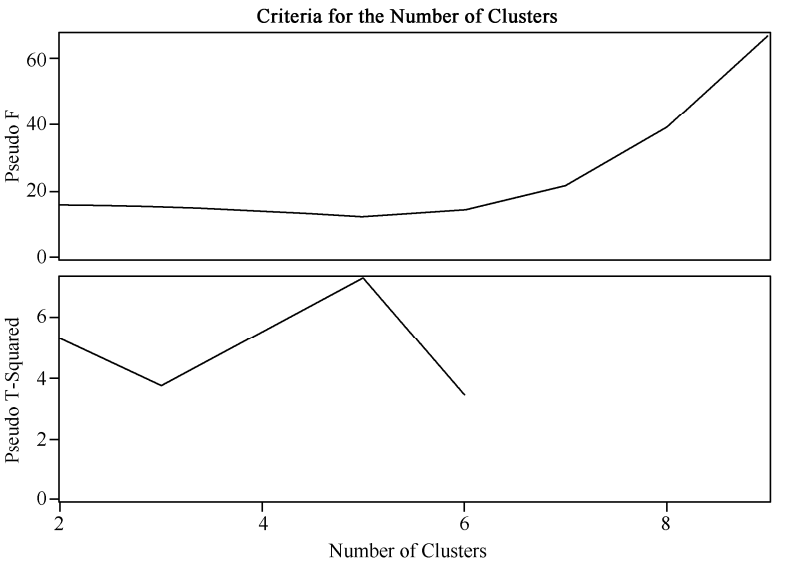


图 12-1　聚类标准图

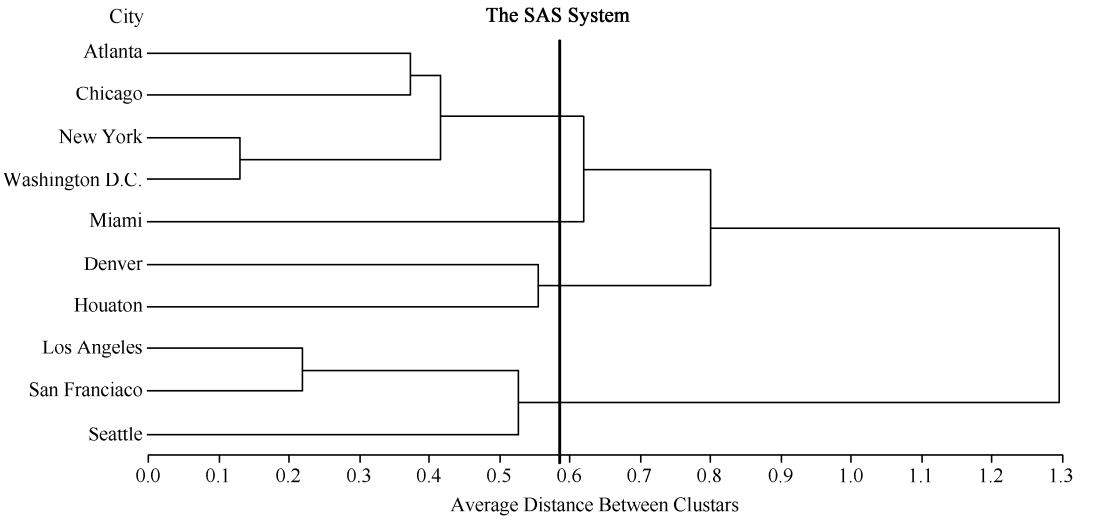


图 12-2　树状图

样品聚类时，不仅可采用类平均法，还可采用最短距离法、最长距离发、重心法、可变类平均法、Ward 离差平方和法和非参数的密度估计法、两阶段密度估计法。在程序语句“method=”后指定不同分析方法。

12.3　变量聚类

12.3.1　SAS 过程——VARCLUSE 过程

VARCLUS 过程基于相关矩阵或协方差矩阵，对数值变量进行不相交或谱系分类。类的划

分通过计算每类第一主成分或重心成分的最大方差而确定，因此，同每一类有联系的是该类中这些变量的线性组合。VARCLUS 过程能够被用来作为变量压缩的方法。对于含有很多变量的变量集，常常用信息损失很少的类分量集替代。若采用相关矩阵的信息，则所有变量都平等；当引用协方差矩阵分析时，某变量有较大方差，该变量则较为重要。VARCLUS 过程生成的输出数据集，可由 SCORE 过程计算出每类的得分。

VARCLUS 过程的一般使用格式如下：

```
PROC VARCLUS      <选项列表>;
VAR          变量列表;
SEED         变量列表;
FREQ         变量;
WEIGHT       变量;
BY           变量列表;
RUN;
```

PROC VARCLUS 语句后的主要控制选项有：控制输入、输出数据集的选项（如表 12-11 所示），控制聚类方法的选项（如表 12-12 所示），控制输出打印的选项等（如表 12-13 所示）。

表 12-11　控制输入、输出数据集的选项

选　　项	意　　义
DATA=SAS 数据集	指定分析数据集，可为原始数据或 TYPE =CORR、UCORR、COV、UCOR、SSCP 或 FACTOR 类型的数据集
OUTSTAT=SAS 数据集	新建一个包含如下统计量的 SAS 数据集：存储均值、标准差、相关系数、类得分系数和聚类结构
OUTTREE=SAS 数据集	新建包含聚类过程的树状结构信息的数据集，供 TREE 过程调用
MINC =N	定义最小聚类个数
MAXC=N	定义最大聚类个数
MAXEIGEN=N	规定每一类中第二特征值所允许的最大值
PERCENT=N	指定类分量必须解释的方差百分比

表 12-12　控制聚类方法的选项

选　　项	意　　义
CENTROID	使用重心成分法聚类
MAXITER= N	规定在交替最小二乘法阶段中的最大迭代次数
MAXSEARCH= N	指定在搜索阶段最大迭代次数
COV	用协方差矩阵聚类
HI	要求在不同层次的类构成谱系聚类结构
INITIAL=	规定初始化类的方法，可选项有 GROUP、INPUT、RANDOM、SEED

表 12-13　控制输出打印的选项

选　　项	意　　义
CORR	打印相关系数矩阵
SIMPLE	打印均值和标准差

续表

选　项	意　义
SHORT	不打印类结构、得分系数和类间相关矩阵
TRACE	列出在迭代过程中每个变量所归入的类
SUMMARY	只打印最后的汇总表

CANDISC 过程中使用的其他语句含义如下：

VAR 语句——指定进行聚类分析的变量。

FREQ 语句——指定频数变量。

WEIGHT 语句——指定加权变量。

BY 语句——指定分组变量。

12.3.2　SAS 实例——对 8 个身体素质指标进行聚类

例 12-2　已知 8 个身体素质指标：臂跨度（ArmSpan）、前臂长度（Forearm）、较短的腿长（LowerLeg）、二转子直径（BitDiam）、胸围（Girth）、胸宽（Width）、身高（Height）和体重（Weight），变量的相关系数矩阵如表 12-14 所示（相应的 SAS 数据集在光盘中的存储路径为“data\chap12\phys8”）。试对这 8 个变量进行聚类分析。

表 12-14　8 个身体素质指标相关系数矩阵

变　量	Height	Weight	ArmSpan	Forearm	LowerLeg	BitDiam	Girth	Width
Height	1	0.473	0.846	0.805	0.859	0.398	0.301	0.382
ArmSpan	0.846	0.376	1	0.881	0.826	0.326	0.277	0.415
Forearm	0.805	0.38	0.881	1	0.801	0.319	0.237	0.345
LowerLeg	0.859	0.436	0.826	0.801	1	0.329	0.327	0.365
Weight	0.473	1	0.376	0.38	0.436	0.762	0.73	0.629
BitDiam	0.398	0.762	0.326	0.319	0.329	1	0.583	0.577
Girth	0.301	0.73	0.277	0.237	0.327	0.583	1	0.539
Width	0.382	0.629	0.415	0.345	0.365	0.577	0.539	1

编程法：

编写如下程序（其在光盘中的存储路径为“proc\chap12\phy8”）：

```
proc varclus data=chap12.phys8   outtree=tree maxc=4;
/*调用 varclus 过程,输出用于绘制树状图的数据集 tree*/
run;
proc tree   data=tree horizontal ;   /*调用 tree 过程绘制水平树状图*/
height _propor_;                     /*定义树形高度为解释的方差率*/
id _name_;                           /*定义变量名为标识*/
run;
```

本例设定最大分类数为 4 类，因为考虑到变量总数为 8，分类个数太多就失去了聚类的本意，若分类个数太少则会损失过多的信息。因此类的选择应该结合专业知识和实际情况确定。

表 12-15 为用间接重心分量聚类分析法得到的第一步，把全部的 8 个变量聚成一类，能解释的方差（Variation Explained）为 4.67288，占总方差比率（Proportion Explained）的 58.41%，并输出以下语句：“Cluster 1 will be split because it has the largest second eigenvalue，1.770983，which is greater than the MAXEIGEN=0 value.”第一类将被分裂，因为它的第二特征根的值为 1.770983410，大于限定的最大特征根 0）。

表 12-15　聚类第一步

Cluster Summary for 1 Cluster					
Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	8	8	4.67288	0.5841	1.7710

表 12-16 为一类分裂成两类的信息，其中第一类包含 4 个变量，能解释的方差为 3.509218，占类中总方差的 87.73%，即 3.509218/4=87.73%；第二类包含 4 个变量，能解释的方差为 2.917284，占类中总方差 72.93%。

表 12-16　聚类第二步

Cluster Summary for 2 Clusters					
Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	4	4	3.509218	0.8773	0.2361
2	4	4	2.917284	0.7293	0.4764

表 12-17 为两类的具体信息，第一类包含变量臂跨度、前臂长、较短腿长和身高，第二类代表二转子直径、胸宽、胸围和体重。R-squared with Own Cluster 列代表的是每一个变量和所在的类的相关系数的平方，R-squared with Next Closest 列代表的是每一个变量和相邻列的相关系数的平方。（1−R**2 Ratio）列的值由同一行的值求得，等于（1−R-squared with Own Cluster）/（1−R-squared with Next Closest），此值越小分类越合理。理想的分类将使组内相关系数的平方值较大，而组间相关系数的平方值较小。如变量 Width 和其所在的 Cluster1 的相关系数的平方值为 0.6329，而和其相邻列 Cluster2 的相关系数的平方值为 0.1619，其对应的 1−R**2 Ratio=(1−0.6329)/(1−0.1619)=0.4380。注意到 Cluster1 中变量对应的 1−R**2 Ratio 的值都较小，而 Cluster2 中变量对应的 1−R**2 Ratio 的值都较大，因此分成两类不太合适。

表 12-17　两类具体信息

2 Clusters		R-squared with		1−R**2 Ratio	Variable Label
Cluster	Variable	Own Cluster	Next Closest		
Cluster 1	ArmSpan	0.9002	0.1658	0.1196	臂跨度
	Forearm	0.8661	0.1413	0.1560	前臂长
	LowerLeg	0.8652	0.1829	0.1650	较短腿长
	Height	0.8777	0.2088	0.1545	身高

续表

2 Clusters		R-squared with		1−R**2 Ratio	Variable Label
Cluster	Variable	Own Cluster	Next Closest		
Cluster 2	BitDiam	0.7386	0.1341	0.3019	二转子直径
	Girth	0.6981	0.0929	0.3328	胸围
	Width	0.6329	0.1619	0.4380	胸宽
	Weight	0.8477	0.1974	0.1898	体重

表 12-18 为分裂成三类的信息，其中第一类包含 4 个变量，能解释的方差为 3.509218，占类中总方差的 87.73%，第二类包含 3 个变量，能解释的方差为 2.386129，占类中总方差的 79.54%，第三类只包含了一个变量，能解释全部的类中方差。表 12-20 为分裂成四类的信息，Cluster1 至 Cluster4 分别能解释类中方差的 87.73%、88.1%、100%和 100%。同时观察表 12-19 和表 12-21，发现聚合成 4 类每个类对应的（1−R**2 Ratio）列的值要小于聚合成三类的值。综上，选择聚合成 4 类，树状图如图 12-3 所示。

注意：由于聚合成三类时，Cluster1 至 Cluster3 能解释总方差的 86.19%，较充分地代表了原始数据的信息，因此本例也可选择聚合成三类。而聚合成 4 类时，能解释总方差的 90.89%，虽然包含了更多原始数据的信息，但是增加了聚类复杂度。在实际的分析中，希望大家能够根据分析背景决定分类个数。

表 12-18　聚类第三步

Cluster Summary for 3 Clusters					
Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	4	4	3.509218	0.8773	0.2361
2	3	3	2.386129	0.7954	0.4184
3	1	1	1	1.0000	

表 12-19　聚成三类对应的统计量

3 Clusters		R-squared with		1−R**2 Ratio	Variable Label
Cluster	Variable	Own Cluster	Next Closest		
Cluster 1	ArmSpan	0.9002	0.1722	0.1205	臂跨度
	Forearm	0.8661	0.1237	0.1529	前臂长
	LowerLeg	0.8652	0.1680	0.1621	较短腿长
	Height	0.8777	0.1939	0.1517	身高
Cluster 2	BitDiam	0.7686	0.3329	0.3469	二转子直径
	Girth	0.7421	0.2905	0.3634	胸围
	Weight	0.8754	0.3956	0.2061	体重
Cluster 3	Width	1.0000	0.4267	0.0000	胸宽

表 12-20　聚类第四步

Cluster Summary for 4 Clusters					
Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	4	4	3.509218	0.8773	0.2361
2	2	2	1.762	0.8810	0.2380
3	1	1	1	1.0000	
4	1	1	1	1.0000	

表 12-21　聚成 4 类对应的统计量

4 Clusters		R-squared with		1−R**2 Ratio	Variable Label
Cluster	Variable	Own Cluster	Next Closest		
Cluster 1	ArmSpan	0.9002	0.1722	0.1205	臂跨度
	Forearm	0.8661	0.1386	0.1555	前臂长
	LowerLeg	0.8652	0.1661	0.1617	较短腿长
	Height	0.8777	0.2153	0.1558	身高
Cluster 2	BitDiam	0.8810	0.3399	0.1803	二转子直径
	Weight	0.8810	0.5329	0.2548	体重
Cluster 3	Width	1.0000	0.4127	0.0000	胸宽
Cluster 4	Girth	1.0000	0.4892	0.0000	胸围

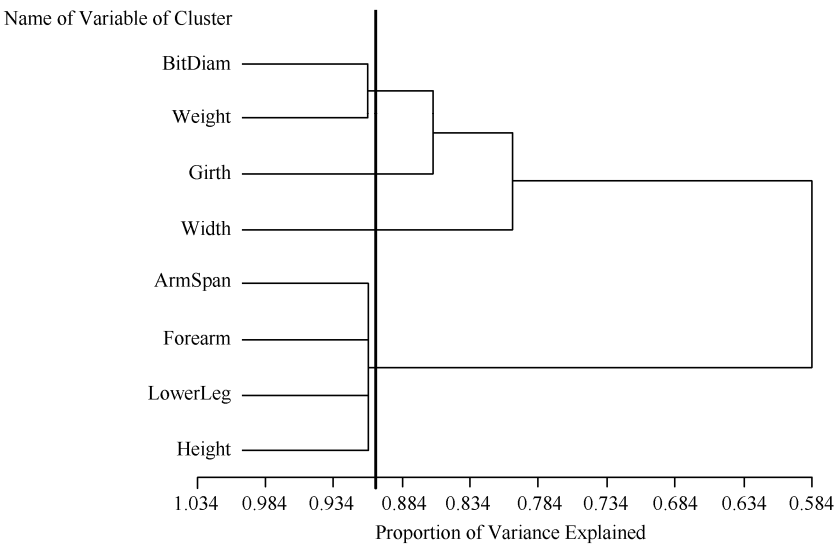


图 12-3　树状图

表 12-22 为标准化变量预测类成分的标准回归系数，据此可得到每一类和变量之间的线性表达式：

Cluster1=0.27038* ArmSpan+0.26519* Forearm+0.26506* LowerLeg+0.26698* Height

Cluster2= 0.53270*BitDiam+0.53270* Weight

Cluster3= Width

Cluster4= Girth

表 12-22　标准化变量预测类成分的标准回归系数（4 类）

Standardized Scoring Coefficients					
Cluster		1	2	3	4
ArmSpan	臂跨度	0.27038	0.00000	0.00000	0.00000
Forearm	前臂长	0.26519	0.00000	0.00000	0.00000
LowerLeg	较短腿长	0.26506	0.00000	0.00000	0.00000
BitDiam	二转子直径	0.00000	0.53270	0.00000	0.00000
Girth	胸围	0.00000	0.00000	0.00000	1.00000
Width	胸宽	0.00000	0.00000	1.00000	0.00000
Height	身高	0.26698	0.00000	0.00000	0.00000
Weight	体重	0.00000	0.53270	0.00000	0.00000

表 12-23 为类结构信息。表 12-24 为类成分之间的相关系数，即 Cluster1 和 Cluster2 之间的相关系数为 0.43174。

表 12-23　类结构信息（4 类）

Cluster Structure					
Cluster		1	2	3	4
ArmSpan	臂跨度	0.94881	0.37395	0.41500	0.27700
Forearm	前臂长	0.93062	0.37236	0.34500	0.23700
LowerLeg	较短腿长	0.93014	0.40751	0.36500	0.32700
BitDiam	二转子直径	0.36620	0.93862	0.57700	0.58300
Girth	胸围	0.30478	0.69943	0.53900	1.00000
Width	胸宽	0.40243	0.64244	1.00000	0.53900
Height	身高	0.93688	0.46398	0.38200	0.30100
Weight	体重	0.44428	0.93862	0.62900	0.73000

表 12-24　类成分间的相关系数

Inter-Cluster Correlations				
Cluster	1	2	3	4
1	1.00000	0.43174	0.40243	0.30478
2	0.43174	1.00000	0.64244	0.69943
3	0.40243	0.64244	1.00000	0.53900
4	0.30478	0.69943	0.53900	1.00000

表 12-25 为聚类分析的综合信息。第二列为对应的分类所能解释的总方差量，第三列为能解释的方差对全部 8 个变量指标的方差百分比，第四列为由一个类所能解释的方差占全部 8 个变量指标的总方差的最小百分比，第五列为各类中最大的第二特征值，第六列为各类中一个变量与其所在类成分的最小相关系数的平方，第七列为各类中$(1-R^2)$Own$-(1-R^2)$Next 的最大比值，即为该次聚类的$(1-R^2)$与最邻近一次聚类的$(1-R^2)$的比值，该值越小说明该次聚类越合理。

表 12-25　聚类综合信息

Number of Clusters	Total Variation Explained by Clusters	Proportion of Variation Explained by Clusters	Minimum Proportion Explained by a Cluster	Maximum Second Eigenvalue in a Cluster	Minimum R-squared for a Variable	Maximum 1−R**2 Ratio for a Variable
1	4.672880	0.5841	0.5841	1.770983	0.3810	
2	6.426502	0.8033	0.7293	0.476418	0.6329	0.4380
3	6.895347	0.8619	0.7954	0.418369	0.7421	0.3634
4	7.271218	0.9089	0.8773	0.238000	0.8652	0.2548

综上所述，本例将变量分成了 4 类，将臂跨度、前臂长、较短腿长和身高归为一类；将二转子直径和体重归为一类；胸围和胸宽各自为一类。

练习题

习题 12-1　已知我国 2009 年各地区农业生产资料价格分类指数，部分数据如表 12-26 所示（包含完整数据的 SAS 数据集在光盘中的存储路径为“data\chap12\farm”）。请据此对各个地区进行聚类分析。

表中变量和对应指标如下：

N1——农用手工工具

N2——饲料

N3——畜产品

N4——半机械化农具

N5——化学肥料

N6——农药及农药械

N7——农用机油

N8——其他农业生产资料

N9——农业生产服务

表 12-26　2009 年城市农业生产资料价格分类指数

city	N1	N2	N3	N4	N5
河　北	104.9	100.5	89.1	100.8	102.2
山　西	101.5	102.4	86.9	99.4	100.4

续表

city	N1	N2	N3	N4	N5
内蒙古	100.9	102.4	89.7	100.8	101.7
辽　宁	102.5	104.7	71.3	100.4	101.4
吉　林	106.7	101.5	86.9	104.9	104.0
city	N6	N7	N8	N9	N10
河　北	101.2	100.7	90.2	100.0	111.2
山　西	99.1	103.5	91.3	102.6	119.7
内蒙古	95.4	98.9	97.3	101.1	105.6
辽　宁	94.1	103.1	87.3	101.9	112.6

（本习题对应的解答程序在光盘中的存储路径为“proc\chap12\farm”。）

习题 12-2　已知条件和习题 12-1 相同，试将这 9 个农业生产资料价格分类指数进行变量聚类。

（本习题对应的解答程序在光盘中的存储路径为“proc\chap12\farm_varcluse”。）

第13章 判 别 分 析

判别分析（Discriminate Analysis）是判别个体所属类别的一种统计方法。它产生于20世纪30年代，近年来，在现代自然科学的许多分支和技术部门中得到广泛的应用。它和上一章介绍的聚类分析的区别在于：聚类分析是按样品的数据特征，把相似的样品倾向于分在同一类中，把不相似的样品倾向于分在不同类中；而判别分析假定类已事先分好，判别新样品应归属哪一类，对类的事先划分常常通过聚类分析得到。

本章首先介绍判别分析的基本原理，再简介SAS系统中应用于判别分析的DISCRIM过程，最后讲解一个判别分析实例。

13.1 基本原理

判别分析方法的分析目的是根据已掌握的一批分类明确的样品，建立一个较准确的判别函数，即使用该判别函数判别时误判率较低，进而能用此判别函数判定新样品所属类别。由此可知，判别分析的关键在于建立判别函数，以及在判定新样品所属类时指定判别规则。以下介绍两种常用的判别分析方法：距离判别分析法和Fisher线性函数判别法。

13.1.1 距离判别分析法

距离判别分析是基于贝叶斯理论的，根据样品 x 属于每一组的先验概率，且在 x 处的可估计的组密度值，由贝叶斯公式计算该样品属于某组的后验概率。设有 k 个组 $G_1,G_2,\cdots,G_k$，且组 G_i 的概率密度为 $f_i(x)$，样品 x 来自组 G_i 的先验概率为 $p_i,i=1,2,\cdots,k$，满足 $\sum_{i=1}^{k}p_i=1$，由贝叶斯公式得到样品 x 属于组 G_i 的后验概率为：

$$p(G_i \mid x)=\frac{p_i f_i(x)}{\sum_{i=1}^{k}p_i f_i(x)}$$

若假设每组内 p 维样品 x 分布为 p 元正态分布情况，即：

$$G_i \sim N_p(\mu_i,\sum_i),(i=1,2,\cdots,k)$$

其中，μ_i 和 $\sum_i$ 分别是第 i 组的均值和协方差矩阵。此时样品 x 来自组 G_i 的概率密度函数为：

$$f_i(x)=(2\pi)^{-p/2}\left|\sum_i\right|^{-1/2}\exp(-0.5d_i^2(x,G_i))$$

其中，$d_i^2(x,G_i)=(x-\mu_i)'\sum_i^{-1}(x-\mu_i)$ 的几何意义为 x 到 i 组均值的平方距离。由此得到样品 x 被归到 G_i 类的后验概率为：

$$p(G_i \mid x)=\frac{p_i \exp(-0.5d_i^2(x,G_i))\left|\sum_i\right|^{-1/2}}{\sum_{i=1}^{k} p_i \exp(-0.5d_i^2(x,G_i))\left|\sum_i\right|^{-1/2}}$$

$$=\frac{\exp(-0.5D_i^2(x,G_i))}{\sum_{i=1}^{k}\exp(-0.5D_i^2(x,G_i))}$$

其中，$D_i^2(x)=d_i^2(x)+g_i+h_i$ 为从样品 x 至第 i 组的广义平方距离。这里

$$g_i=\begin{cases}\log_e\left|\sum_i\right| & \text{若各组协方差矩阵}\sum_i\text{不全相等}\\ 0 & \text{若各组协方差矩阵}\sum_i\text{全相等}\end{cases}$$

$$h_i=\begin{cases}-2\log_e|p_i| & \text{若各组先验概率}\ p_i\ \text{不全相等}\\ 0 & \text{若各组先验概率}\ p_i\ \text{全相等}\end{cases}$$

采用最大后验概率准则判别样品所属的类，即如果样品 x 在第 i 组得到的后验概率 $p(G_i \mid x)$ 为最大值且大于指定的阈值，或者样品 x 到第 i 组的广义平方距离 $D_i^2(x)$ 为最小值，则将样品 x 判归于第 i 组；如果 $p(G_i \mid x)$ 小于指定的阈值，则将样品 x 判归于除 k 组以外的其他组。

介绍了基本判别原理后，接下来将讲解线性判别的判别函数。设有两个协方差矩阵相同的正态总体 G_1 和 G_2，分别服从分布 $N(u_1,\boldsymbol{V})$ 和 $N(u_2,\boldsymbol{V})$，现判定新的样品 y 来自哪个总体。直观上是首先计算样品 y 到两个总体的距离，然后将 y 判至距离较近的一个总体。在实际应用中，计算样品 y 到两个总体的马哈拉诺比斯距离为 $d(y,G_i)=(y-u_i)'\boldsymbol{V}^{-1}(y-u_i)$，则样品至总体 G_1 和 G_2 的距离差为：

$$d(y,G_1)-d(y,G_2)=-2\left(y-\frac{u_1+u_2}{2}\right)'\boldsymbol{V}^{-1}(u_1-u_2)$$

若令

$$w(y)=\left(y-\frac{u_1+u_2}{2}\right)'\boldsymbol{V}^{-1}(u_1-u_2)$$

则判别规则可写成：当 $w(y)\geqslant 0$ 时，$y\in G_1$；当 $w(y)<0$ 时，$y\in G_2$。此时若 u_1、u_2 和 $\boldsymbol{V}$ 已知，则 $w(y)$ 是 y 的线性函数，因此该判别函数被称为线性判别函数。

但是在实际应用中，可能出现两个正态总体协方差矩阵不同的情形，即两个正态总体 G_1 和 G_2 分别服从 $N(u_1,\boldsymbol{V}_1)$ 和 $N(u_2,\boldsymbol{V}_2)$。需要判定新样品 y 来自哪个总体，首先依然选择将样品判定为距离较近的总体，只是样品到两个正态总体的距离差的表达形式发生了变化：

$$d(y,G_1)-d(y,G_2)=y'(\boldsymbol{V}_1^{-1}-\boldsymbol{V}_2^{-1})y-2y'(\boldsymbol{V}_1^{-1}u_1-\boldsymbol{V}_2^{-1}u_2)+u_1'\boldsymbol{V}_1^{-1}u_1-u_2'\boldsymbol{V}_2^{-1}u_2$$

变成了一个二次项判别函数。

一般情况下，即 u_1、u_2 和 $\boldsymbol{V}_1$、$\boldsymbol{V}_2$ 是未知的，假设从两个总体各抽取了 n_1 和 n_2 个样品 $x_1,x_2,\cdots,x_{n_1}$ 和 $y_1,y_2,\cdots,y_{n_2}$。使用线性判别函数还是二次判别函数进行判别分析取决于两个总体的方差。如果两总体方差相等则应用线性判别函数进行判别分析，否则应用二次判别函数判别。于是检验 $\boldsymbol{V}_1$ 与 $\boldsymbol{V}_2$ 是否相等就尤为重要了。

原假设 $H_0:\boldsymbol{V}_1=\boldsymbol{V}_2$，备择假设 $H_1:\boldsymbol{V}_1\neq\boldsymbol{V}_2$

检验统计量：$M=(n_1+n_2-2)\ln\left|\dfrac{S}{n_1+n_2-2}\right|-\sum_{i=1}^{2}(n_i-1)\ln\left|\dfrac{S_i}{n_i-1}\right|$

其中，$\boldsymbol{S}$ 为合并协方差矩阵的估计值，$\boldsymbol{S}_i$ 为第 i 组内的估计协方差矩阵。由于 $(1-d)M$ 近似服从自由度为 f 的卡方 χ^2 分布，其中 $f=p(p+1)/2$，$d=\left(\dfrac{1}{n_1-1}+\dfrac{1}{n_2-1}-\dfrac{1}{n_1+n_2-2}\right)\dfrac{2p^2+3p-1}{6(p+1)}$。

如果 $(1-d)M \geqslant \chi_\alpha^2(p(p+1)/2)$，则在显著性水平 α 下，拒绝原假设 H_0，接受备择假设 H_1；

如果 $(1-d)M < \chi_\alpha^2(p(p+1)/2)$，则在显著性水平 α 下，接受原假设 H_0。

在接受原假设采用线性判别函数时，函数 $w(y)$ 中的 u_1、u_2 和 $\boldsymbol{V}$ 可分别由其无偏估计值代替：

$$w(y)=\left(y-\frac{\bar{u}_1+\bar{u}_2}{2}\right)'\bar{\boldsymbol{V}}^{-1}(\bar{u}_1-\bar{u}_2)$$

其中，

$$\bar{u}_1=\frac{1}{n_1}\sum_{i=1}^{n_1}x_i\ ,\quad \bar{u}_2=\frac{1}{n_2}\sum_{i=1}^{n_2}y_i$$

$$\bar{\boldsymbol{V}}=\frac{1}{n_1+n_2-2}\left[\sum_{i=1}^{n_1}(x_i-\bar{u}_1)(x_i-\bar{u}_1)'+\sum_{i=1}^{n_2}(y_i-\bar{u}_2)(y_i-\bar{u}_2)'\right]$$

在接受备择假设时，使用二次判别函数进行判别分析；二次判别函数中的 u_1、u_2 和 $\boldsymbol{V}_1$、$\boldsymbol{V}_2$ 可分别由其无偏估计值代替：

$$\begin{cases}\bar{u}_1=\dfrac{1}{n_1}\displaystyle\sum_{i=1}^{n_1}x_i\\ \bar{u}_2=\dfrac{1}{n_2}\displaystyle\sum_{i=1}^{n_2}y_i\\ \bar{\boldsymbol{V}}_1=\dfrac{1}{n_1-1}\displaystyle\sum_{i=1}^{n_1}(x_i-\bar{u}_1)(x_i-\bar{u}_1)'\\ \bar{\boldsymbol{V}}_2=\dfrac{1}{n_2-1}\displaystyle\sum_{i=1}^{n_2}(y_i-\bar{u}_2)(y_i-\bar{u}_2)'\end{cases}$$

在实际应用中，可以将两类判别推广到多类判别，判别准则是：将新样品归属于与其距离最短的一个总体类别中。

在判别分析中，有时候会发生误判，即将新样品误判至它本身不属于的总体中，以下讨论两个正态总体的情形下的误判率。协方差相同的两个正态总体 G_1 和 G_2 的分布分别是 $N(u_1,\boldsymbol{V})$ 和 $N(u_2,\boldsymbol{V})$。如果某样品 x 来自 G_1，而且在 $\bar{u}=\dfrac{u_1+u_2}{2}$ 的右边，那么判别规则将判断它来自 G_2，这时就发生了误判（如图 13-1 所示）。

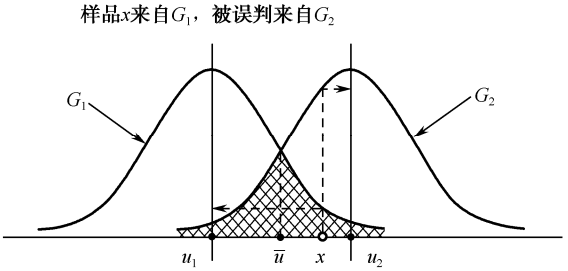


图 13-1　二类判别误判概率图

所谓误判概率的问题是：定义误判概率 P_1 和 P_2，P_1 表示原是第一类的样品而误判为第二类的概率；P_2 表示原是第二类的样品而误判为第一类的概率。误判概率为图中阴影部分的面积。它们为 $P_1=1-\Phi(\overline{u},u_1,\boldsymbol{V})=P_2=\Phi(\overline{u},u_2,\boldsymbol{V})$，这里 Φ 是正态分布的累积分布函数。

13.1.2　Fisher 线性函数判别法

判别分析中寻找合适的判别函数是至关重要的，而在 Fisher 准则下的线性判别函数是一个只要利用总体的一、二阶矩就可求得的比较简单的判别函数。以下以两个总体为例介绍 Fisher 准则下的线性判别函数。

已知一个 p 维向量 $\boldsymbol{x}=(x_1,x_2,\cdots,x_p)'$。设 x_{ijk} 代表第 i 组（$i=1,2,\cdots,r$）中的第 j 个特征（$j=1,2,\cdots,p$）的第 k 个观察值（$k=1,2,\cdots,n_i$）。因此，（$x_{i1k},x_{i2k},\cdots,x_{ipk}$）便相当于第 i 组里面的第 k 个观察所测到的 p 个特征，考虑两个总体，则 $r=2$。现在的目的是找出一个最具有鉴别力的线性判别函数，能比较好地区分这 r 个组，如果 $p>2$，设该线性判别函数为 $\boldsymbol{y}=\boldsymbol{a}'\boldsymbol{x}$，其中 $\boldsymbol{a}'=(a_1,a_2,\cdots,a_p)$。则问题转化成求解系数向量 $\boldsymbol{a}'=(a_1,a_2,\cdots,a_p)$ 使得判别函数 $\boldsymbol{y}=\boldsymbol{a}'\boldsymbol{x}$ 能比较好地区分这 r 个组。

首先将所有的样品值代入线性函数 $\boldsymbol{y}=\boldsymbol{a}'\boldsymbol{x}$ 中得到变量 $\boldsymbol{y}$ 值：$y_{ik}=a_1x_{i1k}+a_2x_{i2k}+\cdots+a_px_{ipk}$，然后考虑所有数据点 y_{ik} 的总变异之和（方差）：$\mathrm{SST}=\sum_{i=1}^{r}\sum_{k=1}^{n_i}(y_{ik}-\overline{y})^2$，其中，$\overline{y}$ 为所有 r 组的总均值，即 $\overline{y}=\boldsymbol{a}'\overline{x}$。对 SST 进行方差的平方和分解成组内方差 SSE 和组间方差 SSR：

$$\begin{aligned}\mathrm{SST}&=\sum_{i=1}^{r}\sum_{k=1}^{n_i}(y_{ik}-\overline{y}_i)^2+\sum_{i=1}^{r}n_i(\overline{y}_i-\overline{y})^2\\&=\mathrm{SSE}+\mathrm{SSR}\end{aligned}$$

其中，$\overline{y}_i$ 表示第 i 组的均值，即 $\overline{y}=\boldsymbol{a}'\overline{x}$，称为组内均值。组间方差 SSR 代表了系统因素引起的变异，而组内方差 SSE 代表了随机因素引起的变异。因此，我们应该选择使得 $\dfrac{\mathrm{SSR}}{\mathrm{SSE}}$ 达到最大值的 a，即表示组与组之间系统因素引起的变异 SSR 比组内随机因素引起的变异 SSE 达到了最大值，使得不同组之间的鉴别力达到最大。

在两个总体时，经过推导得到当且仅当 $a=k(\boldsymbol{V}_1+\boldsymbol{V}_2)^{-1}(u_1-u_2)$ 时，$\dfrac{\mathrm{SSR}}{\mathrm{SSE}}$ 达到最大值 $\dfrac{1}{2}(u_1-u_2)'(\boldsymbol{V}_1+\boldsymbol{V}_2)^{-1}(u_1-u_2)$，其中 k 为任意实数，简单起见可令 $k=1$。于是我们求得的判别函数为 $y=(\boldsymbol{V}_1+\boldsymbol{V}_2)^{-1}(u_1-u_2)x$。

给出判别函数以后，我们还要给出判别准则。取各总体均值的加权平均为比较值，即 $\overline{u}=(\boldsymbol{V}_1+\boldsymbol{V}_2)^{-1}(\boldsymbol{V}_1u_1+\boldsymbol{V}_2u_2)$，相应的样品和总体的距离函数为：

$$D_1=\left\{y:(\boldsymbol{V}_1+\boldsymbol{V}_2)^{-1}(u_1-u_2)(y-\overline{u})\geqslant 0\right\}$$

$$D_2=\left\{y:(\boldsymbol{V}_1+\boldsymbol{V}_2)^{-1}(u_1-u_2)(y-\overline{u})<0\right\}$$

由于总体的特征 u_i 和 $\boldsymbol{V}_i$ 通常是未知的，在应用中可以用样品信息得到它们的无偏估计：判别函数 $y=(\overline{\boldsymbol{V}}_1+\overline{\boldsymbol{V}}_2)^{-1}(\overline{u}_1-\overline{u}_2)x$ 中：

$$\begin{cases} \bar{u}_1 = \dfrac{1}{n_1}\sum_{k=1}^{n_1} y_{1k} \\ \bar{u}_2 = \dfrac{1}{n_2}\sum_{k=1}^{n_2} y_{2k} \\ \bar{V}_1 = \dfrac{1}{n_1 - 1}\sum_{k=1}^{n_1} (y_{1k} - \bar{u}_1)(y_{1k} - \bar{u}_1)' \\ \bar{V}_2 = \dfrac{1}{n_2 - 1}\sum_{k=1}^{n_2} (y_{2k} - \bar{u}_2)(y_{2k} - \bar{u}_2)' \end{cases}$$

令 $\bar{u} = (\bar{V}_1 + \bar{V}_2)^{-1}(\bar{V}_1\bar{u}_1 + \bar{V}_2\bar{u}_2)$，相应的和两个正态总体的距离如下，将点判别到距离较近的总体中。

$$\begin{cases} D_1 = \{y : (\bar{V}_1 + \bar{V}_2)^{-1}(\bar{u}_1 - \bar{u}_2)(y - \bar{u}) \geqslant 0\} \\ D_2 = \{y : (\bar{V}_1 + \bar{V}_2)^{-1}(\bar{u}_1 - \bar{u}_2)(y - \bar{u}) < 0\} \end{cases}$$

13.2 SAS 过程——DISCRIM 过程

在 SAS 系统中主要用 DISCRIM 过程进行判别分析，该过程的一般使用格式如下：

```
PROC DISCRIM   <选项列表> ;
CLASS          变量名;
BY             变量列表;
FREQ           变量名;
ID             变量名;
PRIORS         概率列表;
TESTCLASS      变量名;
TESTFREQ       变量名;
TESTID         变量名;
VAR            变量列表;
WEIGHT         变量名;
RUN;
```

PROC DISCRIM 语句后主要的控制选项分成三类：控制输入、输出数据集选项（如表 13-1 所示）、控制判别分析的类型和规则的选项（如表 13-2 所示）、指定非参数法的选项（如表 13-3 所示）。

表 13-1　控制输入、输出数据集选项

选　项	意　义
DATA=SAS 数据集	指定进行分析的数据集，可为一般 SAS 数据集或几种特殊结构的数据集（协方差矩阵、相关系数矩阵等）
TESTDATA=SAS 数据集	指定欲分类观测的一般 SAS 数据集。注意该数据集中定量变量的变量名必须与 DATA 语句指定的数据集中的变量名匹配。当指定 TESTDATA 选项时，TESTCLASS、TESTFREQ 和 TESTID 语句可用。当采用 TESTDATA 时，输出数据集选项 TESTOUT 和 TESTOUTD 可用来产生检验数据集中观测的分类结果和组密度估计

续表

选　　项	意　　义
OUT=SAS 数据集	生成一个包括输入数据、后验概率和每个观测通过重替换被分入的类等信息的输出数据集
OUTSTAT=SAS 数据集	生成一个包含均值、标准偏差、相关系数和判别统计量等统计量的输出数据集
OUTCROSS=SAS 数据集	生成一个包括来自 DATA 指定数据集的所有数据、后验概率和每个观测通过交叉确认被分入的类的输出 SAS 数据集
OUTD=SAS 数据集	生成一个包含输入数据和每一观测的组密度估计的输出 SAS 数据集
TESTOUT=SAS 数据集	生成一个包含来自 TESTDATA 指定数据集的所有数据、后验概率和每个观测被分入的类的输出 SAS 数据集
TESTOUTD=数据集名	生成一个包含来自 TESTDATA 指定数据集的所有数据和对每一观测的组密度估计的输出 SAS 数据集

表 13-2　控制判别分析的类型和规则的选项

选　　项	意　　义
METHOD=	确定导出分类准则的方法，默认值为 METHOD=NORMAL。当指定 METHOD=NORMAL 时，基于类别内服从多元正态分布，并导出线性或二次判别函数；当指定 METHOD=NPAR 时，采用非参数方法
POOL=	确定广义平方距离的计算是以合并协方差阵还是组内协方差阵为基础。当 POOL=YES 时，采用合并协方差阵得出线性判别函数；当POOL=NO时，采用组内协方差阵得出二次判别函数；当METHOD=NORMAL时，POOL=TEST 要求对组内协方差阵的齐性的似然比检验进行 BARTLETT 修正。默认值为 POOL= YES
SLPOOL= P	当使用控制项 POOL=TEST 时，指定齐性检验的显著性水平。若 POOL=TEST 而 SLPOOL=未指定，系统默认显著性水平为 0.1
THRESHOLD= P	指定分类中可以接受的最小后验概率 *P* 值。若某观察样品归属于某一组的最大后验概率值小于此 *P* 值，那么这个观察样品归入 OTHER 组（已知组以外的组）。系统默认 *P*=0
ANOVA	对各类的单个变量均值之间一元方差分析以此检验判别函数的判别效果
MANOVA	要求对各类的多个变量的均值向量之间进行多元方差分析
LISTERR	要求仅输出由后验概率产生错误分类的样品点相关信息
CROSSLISTERR	要求以交叉表的形式输出实际类别与分类结果之间一致和不一致的信息

表 13-3　指定非参数法的选项

选　　项	意　　义
K=数值	为 K 最近邻规则指定一个 K 值。基于 *X* 的 K 个最近邻得到的信息将观测 *X* 分入一个组
R=数值	为核密度估计指定一个半径 R 值（K=和 R=不能同时设置）
KERNEL	为估计组密度指定一个核密度，默认值为 UNI，此选项与 R=同时设置。可选项有 BIW、EPA、NOR、TRI、UNI
METRIC=	为平方距离的计算指定量度，可选项有 DIAGONAL、FULL、IDENTITY

DISCRIM 过程中使用的其他主要语句含义如下：

VAR 语句——指定进行判别分析的变量子集，并建立关于此变量子集的判别函数。

PRIORS 语句——指定先验概率，它有以下 3 种指定方法：

- PRIORS EQUAL——表示各类先验概率相等，默认值。
- PRIORS PROPORTIONAL——表示各类先验概率等于各类样品频率。
- PRIORS A=P1 B=P2 C=P3——其中 A、B 和 C 是分类标志，P1、P2 和 P3 是先验概率，且 P1+P2+P3=1。

13.3 SAS 实例——根据物质含量判断食物所属类别

例 13-1 已知谷类（Corn）、鲜果类（Fruit）、牲畜类（Livestock）、鱼类（Fish）和豆类（Beans）这 5 类共 100 种食物的 100g 可食部分营养成分表，部分数据如表 13-4 所示（包含完整数据的 SAS 数据集在光盘中的存储路径为“data\chap13\food_train”）。营养成分包括：热量（Heat：单位为千卡）、胆固醇（Cholesterol：单位为克）、蛋白质（Protein：单位为克）、脂肪（Fat：单位为克）、糖类（Sugar：单位为克）、钙（Calcium：单位为微克）和纤维素（Fiber：单位为克）。试根据这些食物建立分类函数并判别表 13-5（相应的 SAS 数据集在光盘中的存储路径为“data\chap13\food_test”）中列出的所属类别，进而得到判别函数的判别准确度。

表 13-4 部分食物的 100g 可食部分主要营养成分表（训练样本）

Type	name	Heat	Cholesterol	Protein	Fat	Sugar	Calcium	Fiber
Corn	白玉米面	340	0	8	4.5	66.9	12	6.2
Corn	标准粉挂面	344	0	10.1	0.7	74.4	14	1.6
Corn	标准粉烙饼	255	0	7.5	2.3	51	20	1.9
Corn	粳米（标四）	346	0	7.5	0.7	77.4	4	0.7
Fruit	菠萝（鲜）	41	0	0.5	0.1	9.5	12	1.3
Fruit	草莓（鲜）	30	0	1	0.2	6	18	1.1
Fruit	橙（鲜）	47	0	0.8	0.2	10.5	20	0.6
Livestock	鸽（鲜）	201	99	16.5	14.2	1.7	30	0
Livestock	公麻鸭（鲜）	360	143	14.3	30.9	6.1	4	0
Livestock	鸡（鲜）	167	106	19.3	9.4	1.3	9	0
Livestock	鸡腿（鲜）	181	162	16.4	13	0	6	0
Livestock	鸡心（鲜）	172	194	15.9	11.8	0.6	54	0
Fish	八爪鱼（鲜）	135	0	18.9	0.4	14	21	0
Fish	鲅鱼（鲜）	122	75	21.2	3.1	2.2	35	0
Fish	白鲢	102	99	17.8	3.6	0	53	0
Beans	扁豆（鲜）	37	0	2.7	0.2	6.1	38	2.1
Beans	菜豆（鲜）	28	0	2	0.4	4.2	42	1.5
Beans	蚕豆（鲜）	104	0	8.8	0.4	16.4	16	3.1
Beans	绿豆芽（鲜）	18	0	2.1	0.1	2.1	9	0.8

表 13-5 26 种食物 100g 可食部分营养成分表（验证样本）

Type	name	Heat	Cholesterol	Protein	Fat	Sugar	Calcium	Fiber
Corn	稻米	346	0	7.4	0.8	77.2	13	0.7
Corn	方便面	472	0	9.5	21.1	60.9	25	0.7
Corn	糯米（优）	344	0	9	1	74.7	8	0.6

续表

Type	name	Heat	Cholesterol	Protein	Fat	Sugar	Calcium	Fiber
Corn	糯米（紫红）	343	0	8.3	1.7	73.7	13	1.4
Fruit	番桃	41	0	1.1	0.4	8.3	13	5.9
Fruit	凤梨	41	0	0.5	0.1	9.5	12	1.3
Fruit	梨（鲜）	32	0	0.4	0.1	7.3	11	2
Livestock	酱鸭	266	107	18.9	18.4	6.3	14	0
Livestock	烤鸡	240	99	22.4	16.7	0.1	25	0
Livestock	鸡翅（鲜）	194	113	17.4	11.8	4.6	8	0
Livestock	猪肉（后蹄膀）	320	145	17	28	0	6	0
Livestock	猪肉（后臀尖）	331	79	14.6	30.8	0	5	0
Fish	鲈鱼（鲜）	100	86	18.6	3.4	0	138	0
Fish	罗非鱼（鲜）	98	78	18.4	1.5	2.8	12	0
Fish	泥鳅（鲜）	96	136	17.9	2	1.7	299	0
Beans	豇豆（鲜）	29	0	2.9	0.3	3.6	27	2.3
Beans	芸豆（鲜）	25	0	0.8	0.1	5.3	88	2.1

编写程序如下所示（其在光盘中的存储路径为“proc\chap13\food”）：

```
proc discrim data=chap13.Food_train                    /*调用 discrim 过程*/
                outstat=foodstat                       /*指定判别函数的输出数据集为 foodstat*/
                method=normal                          /*用参数方法进行判别分析*/
                pool=test                              /*规定检验组内协方差的齐性*/
                list                                   /*列出每个观测的结果*/
                crossvalidate;                         /*进行交叉核实*/
class Type;                                            /*指定分类变量为 Type*/
priors prop;                                           /*按各类的比例计算它们的先验概率*/
id name;                                               /*指定观测的标志变量为 name*/
var Heat Cholesterol Protein Fat Sugar Calcium Fiber ; /*指定分析变量*/
run;

proc discrim data=foodstat   testdata=chap13.food_test
/*调用 discrim 过程；训练样本为 foodstat；检验样本为 food_test*/
testlist;                                              /*列出每个检验样本的结果*/
class Type;                                            /*指定分组变量为 Type*/
testid name;                                           /*指定检验样本的标志变量为 name*/
var Heat Cholesterol Protein Fat Sugar Calcium Fiber ; /*指定分析变量*/
run;
```

选择 Submit|Run 命令提交程序，以下分析主要输出结果。

表 13-6 为样本分类信息，分别列出了 5 个变量的频数、权重和占总体的比例，并将每一列占总体的比例作为它们的先验概率。

表 13-6　样本分类信息

Class Level Information					
Type	Variable Name	Frequency	Weight	Proportion	Prior Probability
Beans	Beans	9	9.0000	0.090000	0.090000
Corn	Corn	21	21.0000	0.210000	0.210000
Fish	Fish	14	14.0000	0.140000	0.140000
Fruit	Fruit	28	28.0000	0.280000	0.280000
Livestock	Livestock	28	28.0000	0.280000	0.280000

表 13-7 第三列为类内协方差的自然对数转换值，对类内斜方差齐性的卡方检验 *P* 值小于 0.001（见表 13-8），即不满足“各类协方差相等”的原假设。在结果输出中有这样一段话：“Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.”（由于在 0.1 的显著性水平之上检验结果显著，则组内斜方差将被用于制定判别函数，参考文献：Morrison, D.F. (1976) Multivariate Statistical Methods p252.）。则本例将用二次判别函数进行判别。

表 13-7　类内协方差矩阵

Within Covariance Matrix Information		
Type	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
Beans	6	−6.26215
Corn	6	5.43334
Fish	6	−2.94947
Fruit	6	3.93626
Livestock	6	9.97718
Pooled	7	28.66497

表 13-8　类内协方差齐性检验结果

Chi-Square	DF	Pr > ChiSq
1876.062208	112	<0.0001

表 13-9 为各类的广义距离平方值。一般情形下，广义距离平方越大的食物类别之间不易误判，反之，广义距离平方值越小的食物类别之间易发生误判。观察此表得到：大豆（Beans）和谷物（Corn）、谷物（Corn）和水果（Fruit）、鱼类（Fish）和牲畜类（Livestock）之间的距离较小，则两两间容易发生误判，而其他食物两两之间的广义距离平方值较大，则不易发生误判。

表 13-9　各类的广义距离平方值

Generalized Squared Distance to Type					
From Type	Beans	Corn	Fish	Fruit	Livestock
Beans	−1.44626	17.52701	1203125705	21.18930	343182616
Corn	3645	8.55463	630640626	29.98268	179885430
Fish	472852837	368676114	0.98276	345286573	23.72736
Fruit	661.25338	12.96687	1075699037	6.48219	306835109
Livestock	638805960	498062886	107.56092	466464612	12.52311

表13-10为应用二次判别函数后对训练样本中每一个观测的判定结果，重新判定的样本如下：将原本属于谷类的小米粥、白萝卜（鲜）归为水果类；将原本属于水果的椰子归为谷类；将原本属于牲畜类的狗肉、牛肉（瘦鲜）、兔肉（鲜）和羊肉（脊背鲜）归于鱼类。

表 13-10　训练样本每个观测的判定情况

Posterior Probability of Membership in Type								
name	From Type	Classified into Type		Beans	Corn	Fish	Fruit	Livestock
白玉米面	Corn	Corn		0.0000	0.9996	0.0000	0.0004	0.0000
标准粉挂面	Corn	Corn		0.0000	1.0000	0.0000	0.0000	0.0000
标准粉烙饼	Corn	Corn		0.0000	0.9999	0.0000	0.0001	0.0000
粳米（标四）	Corn	Corn		0.0000	1.0000	0.0000	0.0000	0.0000
粳米（标一）	Corn	Corn		0.0000	1.0000	0.0000	0.0000	0.0000
粳米（特级）	Corn	Corn		0.0000	1.0000	0.0000	0.0000	0.0000
粳糯	Corn	Corn		0.0000	1.0000	0.0000	0.0000	0.0000
小米粥	Corn	Fruit	*	0.0000	0.0179	0.0000	0.9821	0.0000
燕麦片	Corn	Corn		0.0000	1.0000	0.0000	0.0000	0.0000
薏米	Corn	Corn		0.0000	1.0000	0.0000	0.0000	0.0000
薏苡	Corn	Corn		0.0000	1.0000	0.0000	0.0000	0.0000
油饼	Corn	Corn		0.0000	1.0000	0.0000	0.0000	0.0000
油条	Corn	Corn		0.0000	1.0000	0.0000	0.0000	0.0000
莜麦面	Corn	Corn		0.0000	1.0000	0.0000	0.0000	0.0000
玉米（鲜）	Corn	Corn		0.0000	0.9047	0.0000	0.0953	0.0000
玉米糁（黄）	Corn	Corn		0.0000	0.9997	0.0000	0.0003	0.0000
早籼	Corn	Corn		0.0000	1.0000	0.0000	0.0000	0.0000
早籼（标二）	Corn	Corn		0.0000	1.0000	0.0000	0.0000	0.0000
早籼（标一）	Corn	Corn		0.0000	1.0000	0.0000	0.0000	0.0000
早籼（特等）	Corn	Corn		0.0000	1.0000	0.0000	0.0000	0.0000
白萝卜（鲜）	Corn	Fruit	*	0.4362	0.0702	0.0000	0.4936	0.0000

续表

Posterior Probability of Membership in Type								
name	From Type	Classified into Type		Beans	Corn	Fish	Fruit	Livestock
菠萝（鲜）	Fruit	Fruit		0.0000	0.0159	0.0000	0.9841	0.0000
草莓（鲜）	Fruit	Fruit		0.0001	0.0147	0.0000	0.9853	0.0000
橙（鲜）	Fruit	Fruit		0.0000	0.0249	0.0000	0.9751	0.0000
黄桃（鲜）	Fruit	Fruit		0.0000	0.0466	0.0000	0.9534	0.0000
黄元帅苹果（鲜）	Fruit	Fruit		0.0000	0.0230	0.0000	0.9770	0.0000
金桔（鲜）	Fruit	Fruit		0.0000	0.0076	0.0000	0.9924	0.0000
金枣	Fruit	Fruit		0.0000	0.0076	0.0000	0.9924	0.0000
密云小枣（干）	Fruit	Fruit		0.0000	0.0030	0.0000	0.9970	0.0000
柠檬（鲜）	Fruit	Fruit		0.0006	0.0341	0.0000	0.9654	0.0000
苹果（鲜）	Fruit	Fruit		0.0000	0.0176	0.0000	0.9824	0.0000
苹果梨（鲜）	Fruit	Fruit		0.0000	0.0191	0.0000	0.9809	0.0000
葡萄（鲜）	Fruit	Fruit		0.0000	0.2823	0.0000	0.7177	0.0000
葡萄干	Fruit	Fruit		0.0000	0.0000	0.0000	1.0000	0.0000
青梅果脯	Fruit	Fruit		0.0000	0.0000	0.0000	1.0000	0.0000
人参果（鲜）	Fruit	Fruit		0.0000	0.0932	0.0000	0.9068	0.0000
桑葚（鲜）	Fruit	Fruit		0.0001	0.0274	0.0000	0.9725	0.0000
山楂（鲜）	Fruit	Fruit		0.0000	0.0013	0.0000	0.9987	0.0000
柿（鲜）	Fruit	Fruit		0.0000	0.0198	0.0000	0.9802	0.0000
柿饼	Fruit	Fruit		0.0000	0.0001	0.0000	0.9999	0.0000
桃（鲜）	Fruit	Fruit		0.0000	0.0443	0.0000	0.9557	0.0000
无花果（鲜）	Fruit	Fruit		0.0000	0.0138	0.0000	0.9862	0.0000
香蕉（鲜）	Fruit	Fruit		0.0000	0.0562	0.0000	0.9438	0.0000
杏（鲜）	Fruit	Fruit		0.0000	0.0161	0.0000	0.9839	0.0000
雪花梨（鲜）	Fruit	Fruit		0.0000	0.0186	0.0000	0.9814	0.0000
鸭梨（鲜）	Fruit	Fruit		0.0000	0.0178	0.0000	0.9822	0.0000
椰子	Fruit	Corn	*	0.0000	0.9910	0.0000	0.0090	0.0000
中华猕猴桃（鲜）	Fruit	Fruit		0.0000	0.0154	0.0000	0.9846	0.0000
白瓜子	Fruit	Fruit		0.0000	0.0000	0.0000	1.0000	0.0000
鸽（鲜）	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
公麻鸭（鲜）	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
鸡（鲜）	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
鸡腿（鲜）	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
鸡心（鲜）	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
鸡胸脯肉（鲜）	Livestock	Livestock		0.0000	0.0000	0.0740	0.0000	0.9260
鸡血（鲜）	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000

续表

Posterior Probability of Membership in Type								
name	From Type	Classified into Type		Beans	Corn	Fish	Fruit	Livestock
狗肉	Livestock	Fish	*	0.0000	0.0000	0.9991	0.0000	0.0009
牛肉（后腿鲜）	Livestock	Livestock		0.0000	0.0000	0.2357	0.0000	0.7643
牛肉（前腱鲜）	Livestock	Livestock		0.0000	0.0000	0.4907	0.0000	0.5093
牛肉（前腿鲜）	Livestock	Livestock		0.0000	0.0000	0.3812	0.0000	0.6188
牛肉（瘦鲜）	Livestock	Fish	*	0.0000	0.0000	0.6885	0.0000	0.3115
软五花	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
松江肠	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
蒜肠	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
兔肉（鲜）	Livestock	Fish	*	0.0000	0.0000	0.5903	0.0000	0.4097
午餐肉	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
香肠	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
羊肉（肥瘦鲜）	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
羊肉（脊背鲜）	Livestock	Fish	*	0.0000	0.0000	0.8732	0.0000	0.1268
猪肉（肋条肉）	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
猪肉（奶脯鲜）	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
猪肉（奶面鲜）	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
猪肉（前蹄膀）	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
猪肉（清蒸）	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
猪肉（瘦鲜）	Livestock	Livestock		0.0000	0.0000	0.0474	0.0000	0.9526
猪肉（腿鲜）	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
猪肉香肠罐头	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
八爪鱼（鲜）	Fish	Fish		0.0000	0.0000	0.9876	0.0000	0.0124
鲅鱼（鲜）	Fish	Fish		0.0000	0.0000	0.9857	0.0000	0.0143
白鲢	Fish	Fish		0.0000	0.0000	0.9998	0.0000	0.0002
比目鱼	Fish	Fish		0.0000	0.0000	1.0000	0.0000	0.0000
草鱼（鲜）	Fish	Fish		0.0000	0.0000	0.9966	0.0000	0.0034
鲳鱼（鲜）	Fish	Fish		0.0000	0.0000	0.9967	0.0000	0.0033
大凤尾鱼	Fish	Fish		0.0000	0.0000	1.0000	0.0000	0.0000
大黄花鱼	Fish	Fish		0.0000	0.0000	0.9999	0.0000	0.0001
青鱼（鲜）	Fish	Fish		0.0000	0.0000	0.9769	0.0000	0.0231
武昌鱼	Fish	Fish		0.0000	0.0000	1.0000	0.0000	0.0000
小凤尾鱼	Fish	Fish		0.0000	0.0000	1.0000	0.0000	0.0000
小黄花鱼	Fish	Fish		0.0000	0.0000	1.0000	0.0000	0.0000
小黄鱼（鲜）	Fish	Fish		0.0000	0.0000	1.0000	0.0000	0.0000
鳕鱼（鲜）	Fish	Fish		0.0000	0.0000	0.9929	0.0000	0.0071

续表

Posterior Probability of Membership in Type							
name	From Type	Classified into Type		Beans	Corn	Fish	Fruit	Livestock
扁豆（鲜）	Beans	Beans		0.9896	0.0007	0.0000	0.0096	0.0000
菜豆（鲜）	Beans	Beans		0.9764	0.0029	0.0000	0.0207	0.0000
蚕豆（鲜）	Beans	Beans		1.0000	0.0000	0.0000	0.0000	0.0000
绿豆芽（鲜）	Beans	Beans		0.9618	0.0005	0.0000	0.0377	0.0000
毛豆（鲜）	Beans	Beans		1.0000	0.0000	0.0000	0.0000	0.0000
青豆	Beans	Beans		1.0000	0.0000	0.0000	0.0000	0.0000
豌豆（鲜）	Beans	Beans		0.9996	0.0004	0.0000	0.0000	0.0000
豌豆苗（鲜）	Beans	Beans		0.9897	0.0009	0.0000	0.0094	0.0000
芸豆（鲜）	Beans	Beans		0.9958	0.0000	0.0000	0.0042	0.0000

表 13-11 为应用建立的二次判别函数重新判定样本后的频数统计表。如谷类（Corn）正确判别的样本数为 19 个，占总体的 90.48%，而误判为水果（Fruit）的样本数为 2 个，占总体的 9.52%。其他作物的信息请读者按此方法自行分析。

表 13-11　重新判定样本后的频数统计表

Number of Observations and Percent Classified into Type						
From Type	Beans	Corn	Fish	Fruit	Livestock	Total
Beans	9 100.00	0 0.00	0 0.00	0 0.00	0 0.00	9 100.00
Corn	0 0.00	19 90.48	0 0.00	2 9.52	0 0.00	21 100.00
Fish	0 0.00	0 0.00	14 100.00	0 0.00	0 0.00	14 100.00
Fruit	0 0.00	1 3.57	0 0.00	27 96.43	0 0.00	28 100.00
Livestock	0 0.00	0 0.00	4 14.29	0 0.00	24 85.71	28 100.00
Total	9 9.00	20 20.00	18 18.00	29 29.00	24 24.00	100 100.00
Priors	0.09	0.21	0.14	0.28	0.28	

在表 13-12 中，Rate（比率）行是训练样本的误判率：大豆（Beans）和鱼类（Fish）的误判率为 0；谷物（Corn）的误判率为 9.52%；水果（Fruit）的误判率为 3.57%；牲畜（Livestock）的误判率为 14.29%；总体误判率为 7%。Priors 行是先验概率的值。

表 13-12　训练样本的误判率

Error Count Estimates for Type						
	Beans	Corn	Fish	Fruit	Livestock	Total
Rate	0.0000	0.0952	0.0000	0.0357	0.1429	0.0700
Priors	0.0900	0.2100	0.1400	0.2800	0.2800	

表 13-13 和表 13-14 为对训练样本进行交叉核实判别的结果。交叉核实的基本思想为：为了判断观测 *i* 的判别正确与否，用删除第 *i* 个观测的训练数据样本算出判别函数，然后用此判别函数来判别第 *i* 观测。每一个观测都进行这样的判别。表 13-13 为交叉核实判别的基本情况，其中牲畜（Livestock）被误判为鱼类（Fish）的概率为 17.86%，其他作物误判情况分析类似。表 13-14 为每类作物交叉核实的误判率，其中牲畜（Livestock）的误判率最低，为 17.86%，而大豆（Beans）的误判率达到了 66.67%。

表 13-13　训练样本交叉核实信息

Number of Observations and Percent Classified into Type						
From Type	Beans	Corn	Fish	Fruit	Livestock	Total
Beans	3 33.33	2 22.22	0 0.00	4 44.44	0 0.00	9 100.00
Corn	0 0.00	16 76.19	0 0.00	5 23.81	0 0.00	21 100.00
Fish	0 0.00	0 0.00	10 71.43	0 0.00	4 28.57	14 100.00
Fruit	0 0.00	6 21.43	0 0.00	22 78.57	0 0.00	28 100.00
Livestock	0 0.00	0 0.00	5 17.86	0 0.00	23 82.14	28 100.00
Total	3 3.00	24 24.00	15 15.00	31 31.00	27 27.00	100 100.00
Priors	0.09	0.21	0.14	0.28	0.28	

表 13-14　交叉核实的误判率

Error Count Estimates for Type						
	Beans	Corn	Fish	Fruit	Livestock	Total
Rate	0.6667	0.2381	0.2857	0.2143	0.1786	0.2600
Priors	0.0900	0.2100	0.1400	0.2800	0.2800	

以下分析应用训练样本建立的判别函数作用于测试样本的结果。表 13-15 为应用判别函数对检验样本中所有观测所属类别的判定情况，表的第二列为判定结果，表的第 4 列至第 7 列给出了观测被判为每种作物的对应后验概率，根据贝叶斯判别准则，将观测判定为后验概率最大的一类。19 个观测中只有本属于大豆的新鲜豇豆被判定为了水果。表 13-16 为应用建立的二次

判别函数重新判定样本后的频数统计表，仅大豆的误判率为 50%。表 13-17 为训练样本误判率。

表 13-15　测试样本判别结果

Posterior Probability of Membership in Type								
name	From Type	Classified into Type		Beans	Corn	Fish	Fruit	Livestock
稻米	Corn	Corn		0.0000	1.0000	0.0000	0.0000	0.0000
方便面	Corn	Corn		0.0000	1.0000	0.0000	0.0000	0.0000
糯米（优）	Corn	Corn		0.0000	1.0000	0.0000	0.0000	0.0000
糯米（紫红）	Corn	Corn		0.0000	1.0000	0.0000	0.0000	0.0000
番桃	Fruit	Fruit		0.0000	0.0301	0.0000	0.9699	0.0000
凤梨	Fruit	Fruit		0.0000	0.0159	0.0000	0.9841	0.0000
梨（鲜）	Fruit	Fruit		0.0000	0.0153	0.0000	0.9847	0.0000
酱鸭	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
烤鸡	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
鸡翅（鲜）	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
猪肉（后蹄膀）	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
猪肉（后臀尖）	Livestock	Livestock		0.0000	0.0000	0.0000	0.0000	1.0000
鲈鱼（鲜）	Fish	Fish		0.0000	0.0000	0.5867	0.0000	0.4133
罗非鱼（鲜）	Fish	Fish		0.0000	0.0000	0.9486	0.0000	0.0514
泥鳅（鲜）	Fish	Fish		0.0000	0.0000	1.0000	0.0000	0.0000
豇豆（鲜）	Beans	Fruit	*	0.2407	0.0367	0.0000	0.7226	0.0000
芸豆（鲜）	Beans	Beans		0.9958	0.0000	0.0000	0.0042	0.0000

表 13-16　训练样本判别分析后的频数统计表

Number of Observations and Percent Classified into Type						
From Type	Beans	Corn	Fish	Fruit	Livestock	Total
Beans	1 50.00	0 0.00	0 0.00	1 50.00	0 0.00	2 100.00
Corn	0 0.00	4 100.00	0 0.00	0 0.00	0 0.00	4 100.00
Fish	0 0.00	0 0.00	3 100.00	0 0.00	0 0.00	3 100.00
Fruit	0 0.00	0 0.00	0 0.00	3 100.00	0 0.00	3 100.00
Livestock	0 0.00	0 0.00	0 0.00	0 0.00	5 100.00	5 100.00
Total	1 5.88	4 23.53	3 17.65	4 23.53	5 29.41	17 100.00
Priors	0.09	0.21	0.14	0.28	0.28	

表 13-17　训练样本误判率

Error Count Estimates for Type						
	Beans	Corn	Fish	Fruit	Livestock	Total
Rate	0.5000	0.0000	0.0000	0.0000	0.0000	0.0450
Priors	0.0900	0.2100	0.1400	0.2800	0.2800	

综上所述，本例在对类内协方差矩阵齐性检验发现类内协方差矩阵不等的情况下，系统自动建立了二次判别函数，但由于某些食物之间的广义距离平方值较小，易发生误判，交叉核实结果也验证了这一推论。但是最终根据训练样本给出了需要判别的 19 种食物的类别，仅一种食物发生了误判，说明本判别函数较为有效。

练习题

习题　某科学家采集了 7 种鱼（Bream、Roach、Whitefish、Parkki、Perch、Pike、Smelt）的以下信息：重量（Weight）、高度（Height）、宽度（Width）、从鱼鼻到鱼尾开始处（Length1）、从鱼鼻到鱼尾凹陷处（Length2）、从鱼鼻到鱼尾结束处（Length3），部分数据如表 13-18 所示（包含完整数据的 SAS 数据集在光盘中的存储路径为“data\chap13\fish”）。请根据这些测量指标建立判别函数对鱼进行分类。

表 13-18　鱼类特征指标数据

Species	Weight	Length1	Length2	Length3	Height	Width
Bream	242	23.2	25.4	30	11.52	4.02
Bream	290	24	26.3	31.2	12.48	4.3056
Roach	0	19	20.5	22.8	6.4752	3.3516
Roach	110	19.1	20.8	23.1	6.1677	3.3957
Whitefish	270	24.1	26.5	29.3	8.1454	4.2485
Whitefish	306	25.6	28	30.8	8.778	4.6816
Parkki	140	19	20.7	23.2	8.5376	3.2944
Parkki	170	19	20.7	23.2	9.396	3.4104
Perch	85	18.2	20	21	5.082	2.772
Perch	110	19	21	22.5	5.6925	3.555
Pike	1550	56	60	64	9.6	6.144
Pike	1650	59	63.4	68	10.812	7.48
Smelt	6.7	9.3	9.8	10.8	1.7388	1.0476
Smelt	7.5	10	10.5	11.6	1.972	1.16

（本习题对应的解答程序在光盘中的存储路径为“proc\chap13\fish”。）

反侵权盗版声明